Pontificia Universidad Católica del Ecuador | Sede Ambato

ESCUELA DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

Tema:

DESARROLLO DE UN MODELO DE PREDICCIÓN DE MATRÍCULA PARA UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR

Proyecto de investigación previo a la obtención de título de Ingeniero en Sistemas de la Información

Línea de investigación:

DISEÑO, INFRAESTRUCTURA Y SISTEMAS SOCIALES Y AMBIENTALES PARA UN HÁBITAT SOSTENIBLE

Autor:

Josue Javier Gamboa Cazares

Director:

Mg. Edison Fernando Meneses Torres

Ambato – Ecuador

Julio 2026

DECLARACIÓN DE AUTENTICIDAD Y RESPONSABILIDAD

Yo: **JOSUE JAVIER GAMBOA CAZARES,** con **CC. 185001762-3**, autor del trabajo de graduación intitulado: "DESARROLLO DE UN MODELO DE PREDICCIÓN DE MATRÍCULA PARA UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR", previo a la obtención del título profesional de **INGENIERO EN SISTEMAS DE LA INFORMACIÓN**, en la **ESCUELA DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD.**

1. Declaro tener pleno conocimiento de la obligación que tiene la Pontificia Universidad Católica del Ecuador, de conformidad con el artículo 144 de la Ley Orgánica de Educación Superior, De entregar a la SENESCYT en formato digital una copia del referido trabajo de graduación para que sea integrado.
2. Autorizo a la Pontificia Universidad Católica del Ecuador a difundir a través de sitio web de la Biblioteca de la PUCE Ambato, el referido trabajo de graduación, respetando las políticas de propiedad intelectual de la Universidad.

Ambato, Julio 2026

JOSUE JAVIER GAMBOA CAZARES
CC. 185001762-3

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR SEDE AMBATO

APROBACÍÓN DEL TRIBUNAL DE GRADO

Tema:

DESARROLLO DE UN MODELO DE PREDICCIÓN DE MATRÍCULA PARA UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR

Líneas de Investigación:

DISEÑO, INFRAESTRUCTURA Y SISTEMAS SOCIALES Y AMBIENTALES PARA UN HÁBITAT SOSTENIBLE

Autor:

Josue Javier Gamboa Cazares

Edison Fernando Meneses Torres, Ing. Mg. f_____

1714355482

CALIFICADOR

Verónica Maribel Pailiacho Mena Ing. Mg. f_____

CALIFICADOR

Ricardo Patricio Medina Chicaiza, Ing. PhD. f_____

CALIFICADOR

Francisco Javier Echeverria Tamayo, Ing. Mg. f_____

DIRECTOR DE LA ESCUELA DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

Diego Gonzalo Coca Chanalata, Dr. Mg. f_____

PROSECRETARIO PUCESA

PUCE | AMBATO
PROSECRETARÍA

Ambato-Ecuador

Julio 2026

DEDICATORIA

Este trabajo de titulación se la dedico a Dios que siempre ha estado y estuvo en los momentos que más necesite de alguien, por darme unos maravillosos padres que, aunque no sea el mejor hijo ellos siempre me han apoyado todo, por darme una hermana que siempre estuvo y siempre ha estado, por darme una maravillosa familia la cual siempre me respaldo, a todas las personas que al transcurso de mi vida estudiantil estuvieron.

AGRADECIMIENTO

Quiero agradecer profundamente a Dios por permitirme estar aquí, y a todas las personas, amigos o profesores que se transformaron en amigos y estuvieron en diferentes etapas de mi vida estudiantil, quiero agradecer a la Universidad por permitirme encontrarme cuando pensé que estaba perdido.

RESUMEN

La planificación académica en las instituciones de educación superior depende del número previsto de estudiantes inscritos por periodo y carrera; en la Pontificia Universidad Católica del Ecuador Sede Ambato (PUCESA) esta proyección se realiza con métodos empíricos, lo que limita la asignación de carga docente, la apertura de paralelos y el uso de la infraestructura. El objetivo del presente trabajo fue desarrollar un modelo de predicción de matrícula a partir de los registros del sistema BANNER correspondientes a veintidós periodos académicos consecutivos (2015-P1 a 2025-P2).

El estudio adoptó un enfoque cuantitativo, no experimental y longitudinal retrospectivo, estructurado sobre las seis fases del estándar CRISP-DM; mediante una comparativa empírica validada con un esquema walk-forward se seleccionaron Holt-Winters, regresión de Huber y SARIMA, combinados por mediana, mientras que Random Forest y XGBoost se descartaron por su incapacidad de extrapolación. El ensamble alcanzó un error porcentual absoluto medio (MAPE) de 5,93 % en validación walk-forward y de 5,31 % en validación fuera de muestra frente al periodo 2026-P1, con una proyección de 3.450 estudiantes para 2026-P2 y su intervalo de confianza al 95 %.

Se concluye que el modelo, desagregado por carrera y con clasificación de confiabilidad por niveles, constituye una herramienta interpretable y reproducible que orienta la toma de decisiones institucionales y sienta las bases para su integración en un tablero analítico de planificación académica.

Palabras clave: predicción de matrícula, series temporales, CRISP-DM, planificación académica, educación superior.

ABSTRACT

Academic planning in higher education institutions depends on accurate projections of student enrollment by academic term and academic program. At the Pontificia Universidad Católica del Ecuador Sede Ambato (PUCESA), these projections are currently based on judgment-based methods, which limit the effective allocation of teaching workloads, the opening of course sections, and the efficient use of institutional infrastructure. The objective of this study was to develop an enrollment forecasting model using records extracted from the BANNER system covering twenty-two consecutive academic terms (2015-P1 to 2025-P2). The study adopted a quantitative, non-experimental, retrospective longitudinal design structured around the six phases of the CRISP-DM methodology. Through an empirical comparison validated using a walk-forward approach, the Holt–Winters, Huber regression, and SARIMA models were selected and combined using the median, whereas Random Forest and XGBoost were excluded because of their limited extrapolation capability. The resulting ensemble achieved a mean absolute percentage error (MAPE) of 5.93 % under walk-forward validation and 5.31 % in out-of-sample validation using the 2026-P1 academic term, producing a forecast of 3,450 enrolled students for the 2026-P2 term together with its 95 % confidence interval. The findings indicate that the proposed model, disaggregated by academic program and incorporating a multi-level reliability classification, provides an interpretable and reproducible decision-support tool for institutional planning. Furthermore, it establishes the foundation for integration into an analytical dashboard to support academic planning.

Keywords: *enrollment prediction, time series, CRISP-DM, academic planning, higher education.*

ÍNDICE GENERAL DE CONTENIDOS

INTRODUCCIÓN

Toda institución de educación superior necesita planificar con anticipación sus recursos humanos, físicos y financieros, porque el número de estudiantes matriculados condiciona buena parte de la organización académica y administrativa (Barredo Arrieta et al., 2020). La matrícula se convierte así en un indicador central para decidir sobre la carga docente, la apertura de asignaturas, el uso de la infraestructura y el presupuesto institucional (SENESCYT, 2023).

Sin embargo, la matrícula futura suele estimarse con métodos manuales o con proyecciones basadas en los datos históricos de la institución, que difícilmente capturan los factores académicos, demográficos e institucionales que inciden en el comportamiento de los estudiantes (Sghir et al., 2023). De ahí la necesidad de contar con información anticipada y confiable sobre cuántos estudiantes se matricularán en cada periodo, pues de ese dato dependen la planificación de la carga docente, la organización de la oferta académica y el uso eficiente de los recursos, todo lo cual repercute en la calidad de los procesos educativos.

OBJETIVOS DE LA INVESTIGACIÓN

Objetivo general

Diseñar un modelo de predicción de matrícula para una institución de educación superior.

Objetivos específicos

1. Fundamentar teóricamente la gestión de la matrícula estudiantil y el uso de modelos predictivos en el contexto de la educación superior.
2. Diagnosticar la situación actual de la gestión de la matrícula estudiantil en la institución objeto de estudio.
3. Implementar un modelo de predicción de matrícula mediante técnicas de análisis estadístico y aprendizaje automático.
4. Validar el desempeño del modelo de predicción de matrícula mediante pruebas de caja negra aplicadas a datos históricos.

JUSTIFICACIÓN DE LA INVESTIGACIÓN

La presente investigación se justifica por la necesidad de fortalecer los procesos de planificación académica y administrativa en las instituciones de educación superior, dado que la matrícula estudiantil constituye un factor determinante en la asignación de recursos humanos, físicos y financieros.

Desde el punto de vista institucional, el estudio facilitará la disponibilidad de información anticipada y fiable para tomar decisiones acertadas en la apertura de nuevas materias, la organización de la carga docente y la optimización de la infraestructura académica (Bird, 2023).

La investigación aporta al ámbito académico un análisis de la aplicación de modelos predictivos y técnicas de machine learning en la gestión universitaria y busca potenciar el uso de enfoques cuantitativos y basados en datos dentro de su propia institución (Almalawi et al., 2024). Los resultados que se han obtenido también podrán servir de referencia para futuras investigaciones en analítica institucional, educativa y de planificación.

Desde el punto de vista financiero, la investigación es imperativa, pues asegura la sostenibilidad y optimización económica de las instituciones. Un modelo predictivo fiable de matrícula permite una mejor planificación presupuestaria para los periodos siguientes, disminuyendo el riesgo de sobreestimar ingresos o de incurrir en subejecución por subestimar la demanda. (Perfumo & Ares, 2023).

En el terreno de la gobernanza y el derecho, la investigación trata de ajustarse a los marcos reglamentos vigentes en la educación superior y en la gestión de recursos educativas. Una buena base técnica ayuda a tomar decisiones y apoya procesos importantes —como la apertura o el cierre de programas— con evidencia cuantitativa, lo que reduce los riesgos legales y fortalece la defensa de las decisiones institucionales ante los organismos de control. La última, investigación ayuda a que se cumplan los estándares de calidad educativa, pues una planificación robusta y basada en datos es un requisito implícito de los sistemas de acreditación y evaluación, lo cual ayuda a sostener la estabilidad, la transparencia y la continuidad del servicio educativo.

CAPÍTULO I. ESTADO DEL ARTE Y LA PRÁCTICA

Con la revisión de la literatura aquí presentada se pretende construir la base teórica sobre la que se asienta este trabajo. El capítulo se organiza en torno a tres ejes temáticos, con una continuidad lógica entre ellos: el primero trata la matrícula estudiantil como un indicador estratégico de la gestión universitaria; el segundo sitúa la analítica predictiva en el contexto de la educación superior; y el tercero examina los algoritmos de aprendizaje automático más utilizados para la predicción de series temporales educativas (Sghir et al., 2023). La delimitación de estos ejes está en respuesta al problema identificado en la Pontificia Universidad Católica del Ecuador Sede Ambato (PUCESA) donde la planificación académica aún no cuenta con herramientas formales para poder pronosticar la matrícula. Se consultaron fuentes publicadas en revistas arbitradas, informes de organismos internacionales y estudios empíricos recientes, destacando los trabajos desarrollados en entornos latinoamericanos comparables al contexto institucional analizado (Arias Ortiz et al., 2024).

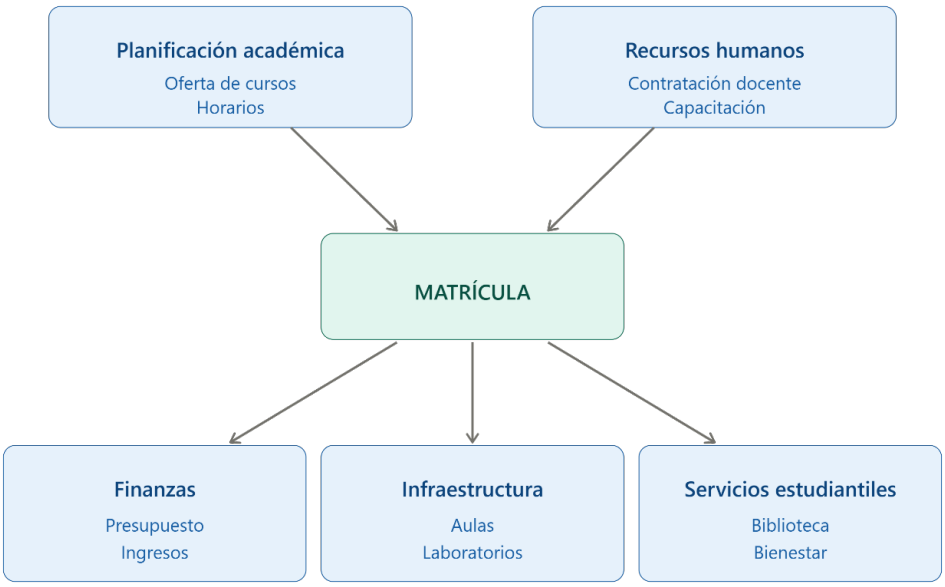
1.1. Gestión de la matrícula estudiantil y educación superior

La educación superior tiene un papel esencial en el desarrollo de la sociedad al ayudar a formar capital humano cualificado, a generar conocimiento y a fortalecer la competitividad de los países (UNESCO, 2021). Las instituciones de educación superior (IES) funcionan en contextos cada vez más dinámicos y competitivos, con cambios demográficos, transformaciones tecnológicas y mayores exigencias de calidad académica. En este contexto, el manejo de la matrícula estudiantil pasa a ser un elemento estratégico para la planificación y la sostenibilidad institucional.

La matrícula estudiantil es el número de estudiantes matriculados en un periodo académico dado y es un indicador importante para la toma de decisiones académicas y administrativas (Perfumo & Ares, 2023). A partir de ella se regula la oferta de asignaturas, la asignación de docentes, el uso de la infraestructura física y tecnológica y la planificación presupuestaria. Al respecto, Fernández (2021) afirma que la matrícula pasa de tener una dimensión estadística a ser el núcleo financiero y académico de la universidad contemporánea; de ella depende tanto la estabilidad

operativa como la viabilidad de los proyectos educativos a mediano y largo plazo. Esta dependencia se muestra en la Figura 1.1.

Figura 1.1. Relación entre la matrícula escolar y los procesos institucionales



Fuente: Adaptado de (Fernández, 2021)

Una estimación incorrecta de la matrícula puede llevar a desequilibrios operativos como la sobrecarga de aulas y docentes o, por el contrario, la subutilización de los recursos institucionales. Estos desajustes tienen un impacto directo sobre la calidad educativa y sobre la salud financiera de las IES (Perfumo & Ares, 2023).

Estos centros han venido manejando la matrícula, tradicionalmente, con análisis de datos históricos muy simples, con proyecciones lineales o con la experiencia de sus responsables académicos. Estos enfoques han permitido una planificación básica, pero presentan limitaciones importantes, no consideran la complejidad de los factores que influyen en el comportamiento de los estudiantes (Bird, 2023). En la Tabla 1.1 se resumen las diferencias entre ambos enfoques.

Tabla 1.1. Comparación entre enfoques tradicionales y modernos en gestión de matrícula

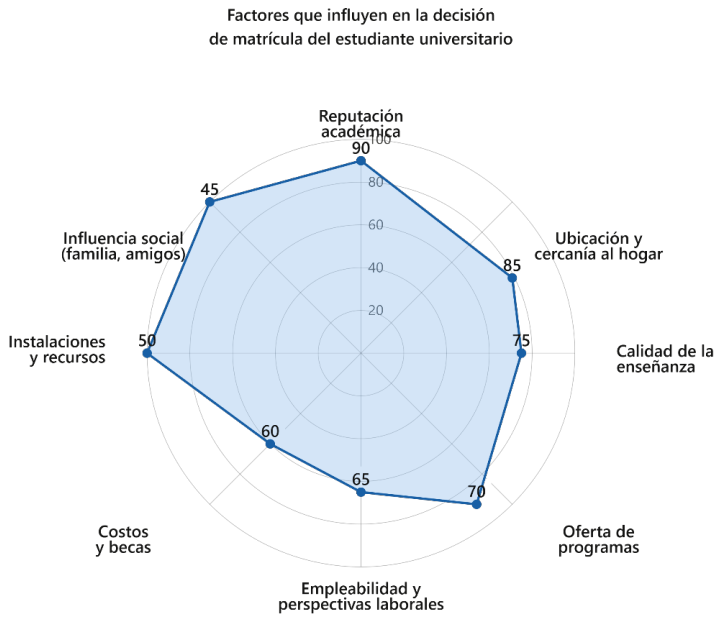
Característica	Enfoque Tradicional	Enfoque Moderno (Predictivo)
Base de decisión	Experiencia institucional, datos históricos simples	Modelos analíticos, múltiples fuentes de datos
Horizonte temporal	Corto plazo (1-2 periodos)	Mediano y largo plazo (3-5 periodos)
Variables consideradas	Limitadas (matrícula histórica)	Múltiples (académicas, demográficas, económicas, sociales)

Característica	Enfoque Tradicional	Enfoque Moderno (Predictivo)
Flexibilidad	Baja, reactiva a cambios	Alta, proactiva ante escenarios
Precisión	Moderada en contextos estables	Alta, incluso en entornos volátiles
Recursos requeridos	Humanos con experiencia	Tecnológicos y humanos especializados

Fuente: Adaptado de (Bird, 2023)

La globalización y la diversificación de la oferta educativa han aumentado la competencia entre las instituciones de educación superior, que se ven obligadas a desarrollar estrategias más precisas para atraer y retener estudiantes. Así, una gestión eficiente de la matrícula no solo influye en lo académico, sino también en la estabilidad financiera y en la reputación institucional (Arias Ortiz et al., 2024). Los principales factores que inciden en esta decisión se resumen en la Figura 1.2.

Figura 1.2. Factores que influyen en la decisión de matrícula del estudiante universitario



Fuente: adaptado de (Briggs, 2006)

Anticipar los cambios en la demanda educativa permite a las universidades responder de manera oportuna a las necesidades del entorno y fortalecer su posicionamiento. Según Fernández (2021), las universidades que han incorporado herramientas de análisis predictivo en su gestión de matrícula han mejorado de forma notable la precisión de su planificación académica y han fortalecido sus indicadores de retención estudiantil.

En países como Colombia y Chile, que han implementado sistemas integrados de gestión estudiantil, se ha observado una relación positiva entre la precisión de las proyecciones de matrícula y los niveles de satisfacción estudiantil (Ministerio de Educación Nacional de Colombia, 2022). Estos sistemas facilitan que:

1. Optimización de los recursos: Distribución proporcional de tutores, equipos y espacios.
2. Personalización de recorridos: Diseño de mallas curriculares ajustadas a perfiles de ingreso.
3. Detección anticipada de riesgos: Identificación de estudiantes con probabilidades de deserción
4. Planificación financiera: Previsión realista de ingresos y gastos operacionales

De ahí que la gestión de la matrícula deba asumir planteamientos que superen los métodos tradicionales. Las herramientas analíticas avanzadas y los modelos predictivos son una alternativa viable para comprender el comportamiento histórico de la matrícula y anticipar los escenarios futuros (Almalawi et al., 2024).

De este modo, las instituciones pueden afianzar sus decisiones estratégicas, optimizar el empleo de los recursos y obtener una planificación académica más eficaz y sostenible.

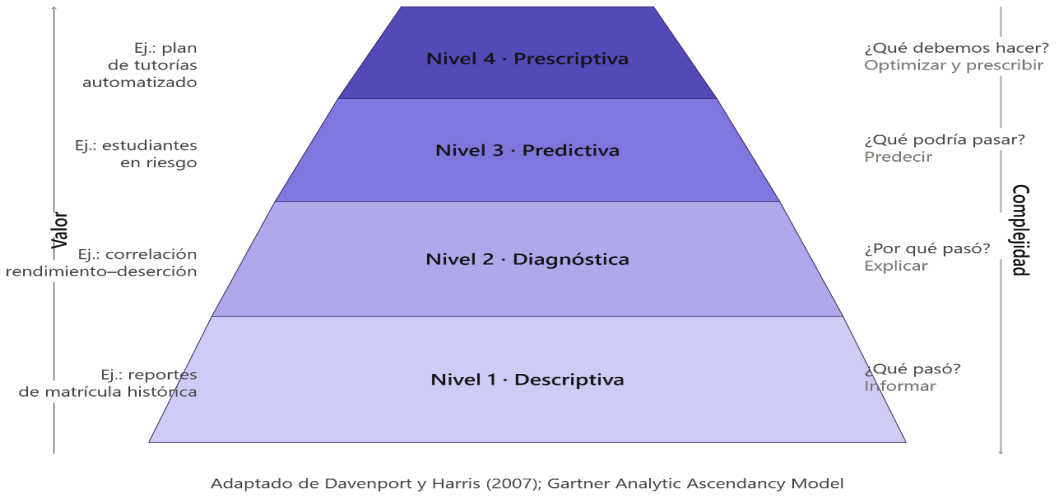
1.2. Analítica predictiva aplicada a la educación superior

La analítica predictiva es una disciplina que integra técnicas estadísticas, métodos computacionales y modelos matemáticos para examinar datos históricos y actuales, con la finalidad de anticipar comportamientos o eventos futuros (Siemens & Long, 2011). En la educación superior ha ganado relevancia debido a la digitalización masiva de los procesos y al creciente volumen de datos que generan los sistemas de información académica, administrativa y financiera de las instituciones.

La analítica educativa no es un concepto único, sino un espectro que crece en complejidad y valor estratégico. Puede entenderse a través de un modelo de madurez

ampliamente aceptado (Campbell & Oblinger, 2007), tal como se presenta en la Figura 1.3:

Figura 1.3. Pirámide de madurez de la analítica en educación superior



Fuente: Adaptado del modelo de Gartner (Gartner, s. f.) *aplicado al contexto educativo*

Aunque las IES se han ubicado tradicionalmente en los niveles 1 y 2, la presión competitiva y la necesidad de eficiencia las empujan hacia los niveles 3 y 4. La analítica predictiva, foco de este estudio, funciona como bisagra en esa transición (Holmes & Tuomi, 2022).

Las instituciones de educación superior disponen actualmente de grandes volúmenes de datos provenientes de múltiples fuentes, muchas veces desconectadas entre sí (Picciano, 2012). El verdadero potencial de la analítica predictiva aparece cuando esos silos de información se integran. La Tabla 1.2 presenta las fuentes más relevantes y su potencial predictivo.

Tabla 1.2. Fuentes de datos y su potencial para la analítica predictiva en IES

Fuente de Datos	Tipo de Datos	Ejemplos de Variables	Aplicación Predictiva Potencial
Sistema de Admisión	Estructurado	Puntaje prueba de ingreso, colegio de procedencia, región.	Predecir probabilidad de matrícula de un admitido.
Sistema Académico (SIGA)	Estructurado	Historial de cursos, calificaciones, asistencia, retiros.	Predecir rendimiento futuro, riesgo de deserción académica.
Plataforma de Aprendizaje (LMS)	Semiestructurado	Tiempo en plataforma, interacciones en foros, descarga de materiales.	Predecir engagement y probabilidad de aprobación.
Sistema Financiero	Estructurado	Estado de pagos, becas, morosidad.	Predecir riesgo de deserción por factores económicos.
Encuestas y Bienestar	Semiestructurado/No estructurado	Satisfacción estudiantil, uso de servicios de salud, participación en actividades.	Predecir bienestar integral y retención.
Datos Externos	Estructurado	Indicadores económicos regionales, tasas de empleo, datos demográficos.	Predecir demanda macro por programas.

Fuente: Elaboración propia basada en (Daniel, 2017)

Sin embargo, Sghir et al. (2023) advierte que contar con grandes volúmenes de datos no garantiza resultados útiles por sí solo; el valor emerge cuando esos datos se integran, limpian y modelan adecuadamente para convertirse en conocimiento accionable que oriente la toma de decisiones.

La analítica predictiva redefine áreas críticas de la gestión universitaria más allá de la matrícula:

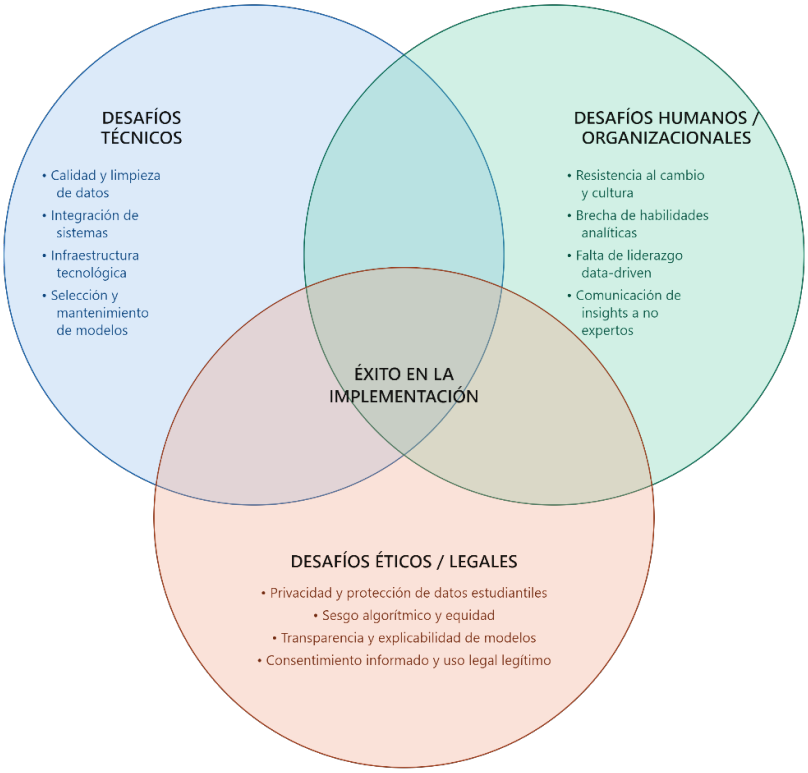
a) Retención y éxito estudiantil: Es el área de aplicación más desarrollada. Modelos como los *Early Alert Systems* analizan combinaciones de variables (rendimiento académico inicial, participación en LMS, datos demográficos) para asignar a cada estudiante un riesgo de deserción.ªl inicio del semestre.

b) Personalización del aprendizaje: Los modelos predictivos pueden recomendar rutas de aprendizaje, recursos adicionales o pares de estudio que se adaptan al estilo y ritmo de cada estudiante. Esto es la base del *Learning Analytics*.

c) Optimización de recursos e infraestructura: Predecir la demanda de cursos específicos permite optimizar la oferta de secciones, la asignación de aulas y la contratación de profesores por horas, lo que mejora la eficiencia operativa y financiera. (Perfumo & Ares, 2023).

La implementación de la analítica predictiva en las instituciones educativas enfrenta importantes desafíos técnicos, humanos y éticos (Slade & Prinsloo, 2013) , como se muestra en la Figura 1.4.

Figura 1.4. Marco de retos para la implementación de analítica predictiva en IES



Fuente: Adaptado de (Siemens & Long, 2011)

El futuro de la analítica predictiva en educación se dirige hacia modelos más transparentes (XAI - Inteligencia Artificial Explicable) que no solamente predigan, sino que expliquen el “porqué” de sus predicciones, generando confianza y permitiendo intervenciones más precisas (Barredo Arrieta et al., 2020).

De la misma manera, surge el concepto de analítica para el bien social, donde la predicción se usa no solo para la eficiencia institucional, sino para identificar y apoyar a poblaciones históricamente marginadas, cerrar brechas de equidad y cumplir con la misión social de la educación superior.(Arias Ortiz et al., 2024).

De esta forma, la analítica predictiva se afirma como herramienta estratégica para las instituciones de educación superior. Si se hace bien, es un paso clave para una gestión universitaria más eficiente, más equitativa y sostenible, siempre y cuando

se haga con rigor técnico, con sentido humano y con ética. (Almalawi et al., 2024; Sghir et al., 2023).

1.3. Aprendizaje automático y modelos predictivos para la matrícula estudiantil

El aprendizaje automático es un campo de la inteligencia artificial que se enfoca en el diseño de algoritmos que sean capaces de reconocer regularidades y realizar predicciones sobre la base de datos pasados, sin que se les indique específicamente cómo hacerlo (Mitchell, 1997). En la educación superior estas técnicas se han establecido como una herramienta útil para el análisis de datos académicos y administrativos, permitiendo modelar comportamientos no lineales y dependencias temporales asociadas a la matrícula estudiantil (Sghir et al., 2023).

En su esencia, la predicción de matrícula es un problema de *forecasting* o predicción de series temporales, donde la variable de interés —el número de inscritos por periodo académico— se estima a partir de su propia trayectoria histórica y, potencialmente, de variables exógenas de relevancia (Almalawi et al., 2024).

Dentro de este enfoque, la presente investigación evalúa de manera empírica un conjunto representativo de algoritmos de tradiciones metodológicas diferenciadas con el objetivo de identificar cuál se adapta de mejor manera a las características específicas de la serie de matrícula institucional: longitud reducida, presencia de tendencia y estacionalidad semestral. En la Tabla 1.3 se detallan los modelos de suavizado exponencial, regresión lineal robusta, modelos autorregresivos estacionales y ensambles basados en árboles que se comparan.

Tabla 1.3. Algoritmos para tener en cuenta para la predicción de matrícula institucional

Algoritmo	Familia metodológica	Principio de funcionamiento
Holt-Winters	Suavizamiento exponencial	Descompone la serie en componentes de nivel, tendencia y estacionalidad, actualizadas de forma recursiva.
Regresión de Huber	Regresión lineal robusta	Estima coeficientes lineales mediante la minimización de una función de pérdida híbrida resistente a valores atípicos.
SARIMA	Modelo autorregresivo estacional (Box-Jenkins)	Integra componentes autorregresivos y de medias móviles con factores estacionales, permitiendo capturar dependencias temporales.
Regresión lineal con lags	Regresión paramétrica supervisada	Relaciona linealmente la matrícula actual con sus valores rezagados y con variables exógenas.
Random Forest	Ensamble por bagging	Promedia múltiples árboles entrenados sobre muestras bootstrap y subconjuntos aleatorios de variables.

Algoritmo	Familia metodológica	Principio de funcionamiento
XGBoost	Ensamble por boosting	Construye árboles secuenciales con regularización incorporada que corrigen los errores residuales del anterior.

Fuente: elaboración propia a partir de Hyndman & Athanasopoulos (2018), Hastie et al. (2009) y Chen & Guestrin (2016).

El análisis parte de la base de la regresión lineal múltiple con variables rezagadas. Su valor está en la transparencia del modelo y en la capacidad de interpretar cada coeficiente directamente como la contribución parcial de la matrícula de los periodos anteriores al valor proyectado (James et al., 2021).

El Random Forest es un método de ensamblado basado en *bagging* que combina las predicciones de un conjunto de árboles de regresión entrenados en muestras *bootstrap* y en selección aleatoria de variables (Hastie et al., 2009). Su mayor ventaja está en disminuir la varianza y en ser robusto ante el sobreajuste, por lo que suele ser una alternativa frecuente en aplicaciones educativas (Bird, 2023). Sin embargo, como está compuesto por árboles cuyas predicciones se restringen a los valores vistos en el entrenamiento, tiene limitaciones para extrapolar tendencias al alza cuando la serie histórica es corta.

XGBoost, en cambio, es una implementación optimizada de *gradient boosting* que entrena árboles en forma secuencial, donde cada árbol nuevo busca reducir el error residual de los anteriores mediante una función objetivo regularizada (Chen & Guestrin, 2016). Esta arquitectura proporciona un rendimiento muy notable en problemas tabulares, especialmente cuando disponemos de un número considerable de observaciones y de variables predictoras.

El paso de construir las variables predictoras, es decir, hacer ingeniería de características (feature engineering), es decisivo para el desempeño de los tres algoritmos. En la predicción de matrícula para el periodo t, por lo general se usan variables rezagadas (matrícula en t-1, t-2 y t-3) para capturar la autocorrelación de la serie, además de indicadores estacionales que separan el primer y el segundo semestre académico (Witten et al., 2017). Se puede complementar con un índice secuencial que refleje la tendencia general del fenómeno.

Para evaluar el desempeño de los modelos predictivos es necesario contar con métricas estadísticamente sólidas y, al mismo tiempo, comprensibles para quienes son

responsables de la planificación académica. En este estudio se utilizan dos métricas complementarias: la raíz del error cuadrático medio (RMSE) expresada en la misma unidad que la variable objetivo, y el error porcentual absoluto medio (MAPE) que expresa el error relativo en términos porcentuales y permite comparar entre carreras de diferente tamaño (Almalawi et al., 2024).

Un aspecto importante para el contexto de este estudio es el largo de la serie histórica disponible. Las series institucionales de matrícula suelen ser cortas, lo que restringe el empleo de arquitecturas profundas y exige la selección de algoritmos cuyo desempeño no se deteriore con conjuntos pequeños de muestras (Almalawi et al., 2024). En tales escenarios, los modelos paramétricos como la regresión lineal con lags pueden ser más convenientes que los métodos basados en árboles, puesto que estos últimos no pueden extrapolar valores más allá del rango observado durante el entrenamiento, mientras que la regresión lineal sí mantiene esta habilidad cuando hay una tendencia sostenida (James et al., 2021).

Por otra parte, la matrícula de una institución de educación superior no se comporta de manera homogénea entre carreras, convivimos con programas con larga trayectoria y matrícula estable, con programas emergentes, programas en declive y programas con interrupciones en su oferta. Esta heterogeneidad recomienda complementar el modelo agregado con un enfoque desagregado por carrera que permita reportar el desempeño del modelo de forma diferenciada y reconocer abiertamente los programas para los cuales la predicción tiene confiabilidad limitada (Perfumo & Ares, 2023).

En conclusión, el aprendizaje automático aporta un conjunto de algoritmos suficientemente sólidos y variados como para encarar el problema de predicción de matrícula. La combinación de un modelo lineal con variables rezagadas, Random Forest y XGBoost permite contrastar enfoques paramétricos y no paramétricos, evaluar tanto el desempeño global como el desempeño por carrera, y obtener predicciones interpretables y útiles para la planificación institucional académica.

CAPÍTULO II. DISEÑO METODOLÓGICO

En este capítulo, se presenta el diseño metodológico en el que se basa el desarrollo del modelo predictivo de matrícula para la Pontificia Universidad Católica del Ecuador Sede Ambato (PUCESA). Se exponen el enfoque y tipo de investigación adoptados, la operacionalización de las variables de estudio, la población y fuente de datos, las técnicas e instrumentos de recolección, el proceso de integración y preparación de la información histórica, y los criterios metodológicos que guían la construcción y evaluación del modelo. La ruta metodológica articula los lineamientos de la investigación cuantitativa con las etapas del estándar CRISP-DM (Cross Industry Standard Process for Data Mining), reconocido como marco de referencia en proyectos aplicados de ciencia de datos (Witten et al., 2017).

2.1. Enfoque y tipo de investigación

El estudio se basa en un enfoque cuantitativo, que se apoya en el análisis numérico de series históricas de matrícula, mediante técnicas estadísticas y de aprendizaje automático supervisado. Esta forma de proceder es acorde con la naturaleza del problema estudiado, el fenómeno observado —número de estudiantes matriculados por programa y periodo académico— es estrictamente medible y admite tratamiento formal a través de modelos matemáticos que permiten cuantificar tendencias, estacionalidades y márgenes de error (Hernández-Sampieri et al., 2022; James et al., 2021).

El diseño de investigación es no experimental, puesto que el investigador no manipula ni interviene en las variables involucradas, sino que observa el fenómeno tal como ocurre en su contexto natural a partir de los registros administrativos ya generados por la institución (Hernández-Sampieri et al., 2022). En el ámbito de esta clasificación, el estudio adopta un diseño longitudinal de tendencia, la información se recoge a lo largo de veintidós periodos académicos consecutivos (2015-P1 a 2025-P2); el periodo 2026-P1 se mantiene para validación fuera de muestra, con el objetivo de observar la evolución del fenómeno en el tiempo y descubrir patrones de comportamiento. Su carácter es además retrospectivo, dado que los datos analiza-

dos corresponden a hechos ya ocurridos y consolidados en el sistema institucional (Brusnahan et al., 2022).

El tipo de investigación es aplicada, puesto que el conocimiento generado no persigue un fin teórico abstracto, sino la resolución de un problema concreto de planificación institucional en PUCESA: la ausencia de un mecanismo formal de pronóstico de matrícula (Bird, 2023). Finalmente, su alcance es descriptivo-predictivo: describe el comportamiento histórico de la matrícula mediante estadística descriptiva y, sobre esa base, infiere o proyecta su valor esperado en periodos futuros mediante modelos de pronóstico (Almalawi et al., 2024).

El desarrollo del modelo se estructura conforme a las seis fases del estándar CRISP-DM, lo que garantiza un proceso ordenado, iterativo y replicable. La Tabla 2.1 sintetiza la correspondencia entre cada fase del estándar y las actividades concretas ejecutadas en la presente investigación.

Tabla 2.1. Correspondencia entre las fases de CRISP-DM y las actividades de la investigación

Fase CRISP-DM	Actividad desarrollada en el estudio
Comprensión del negocio	Delimitación del problema de planificación de matrícula en PUCESA y definición de los objetivos del modelo predictivo.
Comprensión de los datos	Exploración inicial de los reportes BANNER, identificación de campos, formatos y problemas de calidad del dato.
Preparación de los datos	Integración de fuentes heterogéneas (Excel, CSV), reduplicación por PIDM+periodo, normalización de carreras y construcción de variables.
Modelado	Entrenamiento y comparación empírica de los algoritmos candidatos sobre las series de matrícula.
Evaluación	Contraste del desempeño mediante validación walk-forward y métricas de error (MAPE, MAE, RMSE).
Despliegue	Traducción de los resultados en proyecciones y en un tablero interactivo orientado a la planificación institucional.

Fuente: elaboración propia, a partir de Witten et al. (2017).

La operacionalización constituye el proceso mediante el cual las variables conceptuales se traducen en dimensiones e indicadores medibles, y se establece de manera explícita cómo se observará y registrará cada una de ellas (Hernández-Sampieri et al., 2022). De acuerdo con el enfoque cuantitativo y a los objetivos específicos de la investigación, el estudio considera dos variables principales.

La variable dependiente es la matrícula estudiantil por programa académico y periodo, que se operacionaliza mediante el número de estudiantes inscritos por carrera en cada semestre, de acuerdo con los registros del sistema BANNER. Es la variable

objetivo que el modelo trata de estimar. La variable independiente corresponde al modelo predictivo desarrollado, entendido como el conjunto de algoritmos, variables explicativas y procedimientos de validación que generan la proyección, y su desempeño se operacionaliza mediante métricas de error como el error porcentual absoluto medio (MAPE), el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE) para cuantificar la distancia entre los valores predichos y los observados (Almalawi et al., 2024).

Para medir ambas variables se utilizan escalas mixtas, las cuales se seleccionan en función de la naturaleza de cada dato con el fin de preservar la información original de los registros administrativos sin someterlos a transformaciones artificiales (Hernández-Sampieri et al., 2022). Se emplea escala nominal para los identificadores administrativos (códigos BANNER, formato de archivo, modalidad de estudio); escala ordinal para las banderas de calidad del dato (registro completo, periodo COVID, programa nuevo) y para los ítems de percepción institucional; y escala de razón para los conteos cuantitativos (número de estudiantes inscritos y matriculados), por tratarse de magnitudes con cero absoluto sobre las que son válidas las operaciones aritméticas que requiere el modelado.

La Tabla 2.2 presenta la matriz de operacionalización, que articula cada objetivo específico con la variable, la dimensión, los indicadores y los instrumentos que permiten su medición, lo que garantiza la coherencia metodológica entre el problema, los objetivos y la recolección de datos.

Tabla 2.2. Matriz de operacionalización de variables

Variable	Dimensión	Indicadores	Escala	Instrumento
Matrícula estudiantil (dependiente)	Volumen de inscritos por carrera y periodo	N.º de estudiantes inscritos; N.º de matriculados; periodo_id; familia de carrera	Razón / Nominal	Ficha de Análisis Documental (Anexo A)
Modelo predictivo (independiente)	Desempeño del pronóstico	MAPE; MAE; RMSE; tier de confiabilidad	Razón / Ordinal	Ficha de Análisis Documental (Anexo A)
Contexto institucional (de apoyo)	Procesos y calidad del dato	Percepción sobre procesos, calidad, prácticas de proyección y factores exógenos	Ordinal (Likert 1–5)	Guía de Entrevista Semiestructurada (Anexo B)

Fuente: elaboración propia.

Esta operacionalización evidencia que la fase de modelado y validación (Objetivos Específicos 3 y 4) se nutre fundamentalmente de la variable dependiente medida

con la Ficha de Análisis Documental, mientras que la fase de diagnóstico (Objetivo Específico 2) se complementa con la información contextual obtenida mediante la entrevista semiestructurada, lo que conforma así un diseño metodológico integrado y trazable.

2.2. Población y muestra

Población

La población de estudio está conformada por la totalidad de los registros administrativos de matrícula estudiantil de pregrado de la Pontificia Universidad Católica del Ecuador Sede Ambato (PUCESA), almacenados en el sistema institucional BANNER, correspondientes al periodo comprendido entre el primer semestre de 2015 (2015-P1) y el segundo semestre de 2025 (2025-P2); el primer semestre de 2026 (2026-P1) se reserva como periodo de validación fuera de muestra. Se considera periodo P1 al ciclo académico que inicia en febrero (códigos BANNER 61 para grado y 51 para Medicina) y periodo P2 al que inicia en agosto (códigos 66 y 56, respectivamente). En su totalidad, la población abarca veintidós periodos académicos consecutivos y comprende todos los programas de pregrado activos durante este lapso (SENESCYT, 2023).

El nivel de agregación es el número de estudiantes matriculados por programa académico y por periodo. No se trabaja con datos de los estudiantes de forma individual —cumpliendo con las consideraciones éticas detalladas en la sección de validez y confiabilidad—, sino con totales institucionales que permiten la modelación de series temporales sin comprometer la confidencialidad de la información personal (Slade & Prinsloo, 2013).

Muestra

Se procede a utilizar un muestreo no probabilístico de tipo censal o exhaustivo, que incluye todos los registros disponibles para los programas que cumplen con el

criterio de cobertura histórica completa, el objeto de estudio son series temporales agregadas de instituciones y no individuos (Hernández-Sampieri et al., 2022). Como criterio de inclusión metodológica, el programa académico debe poseer datos de los veintidós periodos del horizonte temporal analizado, condición indispensable para el entrenamiento de los modelos de series temporales sin recurrir a imputaciones masivas que comprometan la validez de las predicciones (Hyndman & Athanasopoulos, 2018).

Bajo este criterio, la muestra final queda conformada por seis programas de pregrado de PUCESA, presentados en la Tabla 2.3.

Tabla 2.3. Muestra de programas académicos seleccionados

Programa académico	Escuela / Facultad	Periodos disponibles
Derecho	Ciencias Jurídicas y Sociales	22 (2015-P1 a 2026-P1)
Administración de Empresas	Ciencias Administrativas y Económicas	22 (2015-P1 a 2026-P1)
Contabilidad y Auditoría	Ciencias Administrativas y Económicas	22 (2015-P1 a 2026-P1)
Diseño	Hábitat, Infraestructura y Creatividad	22 (2015-P1 a 2026-P1)
Psicología Clínica	Ciencias de la Salud	22 (2015-P1 a 2026-P1)
Ingeniería en Sistemas / Sistemas de la Información / Tecnologías de la Información	Hábitat, Infraestructura y Creatividad	22 (2015-P1 a 2026-P1)

Fuente: elaboración propia, a partir de los registros del sistema BANNER de PUCESA.

Los programas que iniciaron su oferta académica en periodos posteriores a 2015 —como el caso de Medicina, cuya primera cohorte corresponde a 2021-P2— quedan excluidos de la modelación predictiva por carecer de la profundidad histórica requerida; no obstante, sus registros forman parte del análisis descriptivo institucional. De igual manera, los periodos comprendidos entre 2020-P1 y 2021-P1, marcados por la pandemia de COVID-19, se señalan como anómalos y se controlan mediante una variable indicadora durante el aprendizaje de patrones regulares, aunque se conservan para el análisis exploratorio y las pruebas de robustez (Almalawi et al., 2024).

2.3. Técnicas e instrumentos de recolección de datos

La elección de las técnicas de recolección se corresponde con el enfoque cuantitativo de la investigación, su condición de no experimental, longitudinal y retrospectiva, así como con la naturaleza de la unidad de análisis (registros agregados de matrícula de instituciones). Hernández-Sampieri y cols. (2022) afirman que la elección del instrumento debe responder a la pregunta sobre qué información se necesita y de qué fuentes se obtiene; en el presente estudio, las fuentes primarias son los reportes administrativos del sistema BANNER, mientras que las fuentes secundarias se obtienen mediante la consulta a informantes institucionales que conocen los procesos académico-administrativos.

De ahí la aplicación de dos técnicas complementarias: el análisis documental, como técnica principal por su adecuación con datos cuantitativos preexistentes, y la entrevista semiestructurada, como técnica de apoyo para contextualizar los hallazgos numéricos con información cualitativa institucional. Cada técnica cuenta con un instrumento formal que se describe a continuación, cuyo modelo en blanco se incluye en los Anexos A y B.

El análisis documental es una técnica de investigación que consiste en la revisión sistemática y estructurada de fuentes documentales, ya sean éstas físicas o digitales, primarias o secundarias, con el fin de extraer información relevante para los objetivos del estudio (Hernández-Sampieri et al., 2022). En estudios cuantitativos que utilizan fuentes administrativas, esta técnica es especialmente útil, permitiendo trabajar con datos objetivos, verificables y libres del sesgo introducido por la respuesta de los informantes (Daniel, 2017).

Para el presente trabajo de investigación, se empleó el análisis documental sobre la base de los reportes oficiales del sistema BANNER de PUCESA correspondientes al horizonte 2015-P1 a 2026-P1. Se usan tres tipos de archivos dependiendo del periodo: reportes en Excel multi pestaña para los periodos 2015–2019, archivos CSV de “Estadísticas de Inscritos” para los periodos 2020–2025, y reportes de “Asignaturas por estudiante” para el periodo 2026-P1. Se justifica la elección de esta técnica por tres razones: la disponibilidad efectiva y completa de los registros, la objetividad propia de una fuente sistémica institucional, y la consistencia metodológica con el carácter cuantitativo del estudio.

Instrumento: Ficha de Análisis Documental (FAD)

El instrumento formal asociado es la Ficha de Análisis Documental (FAD), una matriz estructurada que permite homogeneizar la extracción y registro de los datos cuantitativos de los reportes BANNER, lo cual garantiza la trazabilidad y comparabilidad entre periodos. La ficha consta de cuatro apartados: identificación de la fuente, datos cuantitativos extraídos, observaciones sobre la calidad de los datos y validación. El modelo en blanco se muestra en el Anexo A. A continuación se detallan las variables que se registran mediante la FAD:

- Identificación de la fuente: nombre de archivo, formato, periodo BANNER (códigos 61/51/66/56), año, semestre y fecha de extracción.
- Variables cuantitativas: programa académico, número de estudiantes inscritos, número de estudiantes matriculados, modalidad de estudio, código BANNER de la carrera.
- Variables derivadas: identificador secuencial de periodo (periodo_id), familia normalizada de carrera y banderas de calidad (registro completo, periodo COVID, programa nuevo).
- Observaciones cualitativas: faltantes detectados, inconsistencias detectadas, valores atípicos, codificaciones especiales (caso Medicina, repotenciación de mallas, sufijos _R o _Repotenciada) y notas de validación cruzada con totales institucionales.

La entrevista semiestructurada es una técnica de obtención de información basada en la conversación orientada entre el investigador y un informante clave, a través de un conjunto predeterminado de preguntas que permiten la flexibilidad para ahondar en los temas que surjan (Hernández-Sampieri et al., 2022). Aunque la investigación aquí realizada es mayormente cuantitativa, esta técnica complementaria permite contextualizar la información numérica con datos institucionales sobre los procesos de matrícula, las decisiones académicas que pudieron incidir en las series, los cambios curriculares (como las repotenciaciones de mallas) y las prácticas actuales de planificación que el modelo predictivo pretende apoyar.

La elección de esta técnica se justifica por la necesidad de identificar fenómenos institucionales no observables directamente desde los reportes administrativos: por ejemplo, los motivos del cambio en los códigos de carrera, el impacto de campañas

específicas de admisión, los criterios actualmente utilizados para proyectar la matrícula y la identificación de variables exógenas que la institución considere relevantes. Esta información cualitativa, sirve como insumo para la fase de diagnóstico (Objetivo Específico 2), y permite validar la plausibilidad institucional de los resultados del modelo (Objetivo Específico 4).

Instrumento: Guía de Entrevista Semiestructurada (GES)

El instrumento asociado es la Guía de Entrevista Semiestructurada (GES), la cual está organizada en bloques temáticos acordes con los objetivos del estudio: procesos institucionales de matrícula, gestión y calidad de la información, prácticas actuales de planificación y proyección, factores institucionales que inciden en la matrícula y expectativas frente a un modelo predictivo. Los ítems de percepción se midieron mediante una escala de Likert con cinco puntos. El formato completo del instrumento se muestra en el Anexo B.

Con el objetivo de asegurar la validez de contenido y la trazabilidad de la metodología, en la Tabla 2.4 se relaciona cada objetivo específico con la técnica, el instrumento y la dimensión evaluada. La matriz de coherencia permite ver que todos los objetivos tienen un mecanismo definido de recolección de datos y que no hay instrumento que recoja datos que no estén relacionados con los propósitos de la investigación (Hernández-Sampieri et al., 2022).

Tabla 2.4. Coherencia entre objetivos, técnicas, instrumentos y dimensiones medidas

Objetivo específico	Técnica	Instrumento	Dimensión medida
OE2. Diagnosticar la situación actual de la gestión de la matrícula.	Análisis documental y entrevista semiestructurada	FAD (Anexo A) y GES (Anexo B)	Procesos institucionales y calidad del dato
OE3. Implementar un modelo predictivo de matrícula.	Análisis documental cuantitativo	FAD (Anexo A)	Variables cuantitativas BANNER
OE4. Validar el desempeño del modelo (caja negra).	Análisis documental (walk-forward y out-of-sample)	FAD (Anexo A)	Variables derivadas y validación cruzada

Fuente: elaboración propia.

La validez de contenido de ambos instrumentos se asegura mediante su construcción a partir del marco teórico y de la matriz de operacionalización, así como por la revisión y el juicio del director del trabajo de titulación. En el caso de la Ficha de Análisis Documental, la confiabilidad está garantizada por la objetividad y verificabilidad de la fuente: al extraer los datos directamente del sistema BANNER sin mediacio-

nes interpretativas, la medición es reproducible y se valida mediante el contraste cruzado de los totales por periodo con los reportes institucionales oficiales (Daniel, 2017).

En cuanto a la Guía de Entrevista Semiestructurada aplicada al informante experto, la confiabilidad se garantiza mediante la combinación de tres mecanismos coherentes con la modalidad de entrevista a informante clave: la construcción del instrumento a partir del marco teórico y de los objetivos específicos del estudio, que asegura su validez de contenido; la grabación íntegra y transcripción literal de la entrevista, que permite la trazabilidad de los registros; y la validación posterior por parte del propio informante mediante *member checking*, procedimiento estándar para corroborar la interpretación realizada por el investigador (Hernández-Sampieri et al., 2022). El tratamiento de la información se rige por principios de confidencialidad, anonimización y uso exclusivamente académico de los datos institucionales, en concordancia con las consideraciones éticas que enmarcan la investigación (Slade & Prinsloo, 2013).

2.4. Procesamiento y unificación de los datos

En el estándar CRISP-DM la fase que mayor exigencia técnica presenta es la de la preparación de los datos, por lo que de su rigor depende la validez de todo el modelado posterior (Witten et al., 2017). Las fuentes documentales descritas presentan formatos heterogéneos —Excel multi pestaña, archivos CSV de inscritos y reportes de asignaturas por estudiante—, con codificaciones distintas para una misma carrera y con campos no homogéneos entre periodos, por lo que se diseña un proceso de extracción, transformación y carga (ETL) que unifica toda la información en una única estructura analítica coherente.

El proceso de extracción consolida las tres familias de archivos en un único dataset. Los reportes en formato Excel (2015-2019) se leen hoja por hoja, cada hoja corresponde a un periodo académico, los archivos CSV de “Estadísticas de Inscritos” (2020-2025) se leen con delimitador de punto y coma y conservan los campos identificatorios y de programa y los reportes de “Asignaturas por estudiante” (2026-P1) se procesan a nivel de estudiante para derivar el conteo de matriculados por carrera. Los campos PIDM, PERIODO y PROGAMA_DESC funcionan como llaves comunes que posibilitan la integración entre fuentes de manera transversal.

En la fase de limpieza se atienden los problemas de calidad detectados durante la comprensión de los datos. Lo primero que se hace es eliminar los registros duplicados por medio de la combinación única de identificador de estudiante y periodo (PIDM + PERIODO), evitando así el doble conteo que genera la inscripción varias veces a una asignatura dentro de un mismo semestre. En segundo lugar, los registros del año 2020 que presentan el campo TIENE_PAGO con valor “NO” corresponden a preinscripciones no confirmadas y no a matrículas efectivas, por lo que se excluyen del conteo de matriculados, decisión metodológicamente necesaria para no sobreestimar la serie. En tercer lugar, se depuran inconsistencias menores de formato como los espacios en blanco iniciales en el campo género que de no corregirse fragmentarían las categorías durante la agregación.

Un programa académico aparece con diferentes nombres dependiendo de la fuente: en mayúsculas en los reportes Excel, con mayúsculas y minúsculas mezcladas y sufijos _R o _Repo en los CSV de BANNER, y con sufijo _Repotenciada en el formato de asignaturas de 2026. Para hacer frente a esta heterogeneidad, se implementa una función de normalización, denominada familia_carrera(), que realiza un mapeado de todas las variantes de denominación a una única familia de carrera, estable en el tiempo. Esta normalización es indispensable para garantizar que la serie temporal de cada programa sea continua y comparable a lo largo de los veintidós periodos, condición sin la cual el modelado por carrera no sería posible.

En particular, la carrera de Medicina cuenta con sus propios códigos de período BANNER (terminaciones 51 y 56 en vez de 61 y 66), que identifican su calendario alternativo. Esta convención técnica se preserva durante el procesamiento para evitar la pérdida de información de un programa con calendario diferenciado, aun cuando dicho programa quede fuera de la modelación predictiva por su limitada profundidad histórica.

Con el fin de ordenar cronológicamente las observaciones y permitir el cálculo de variables temporales, se construye un identificador secuencial de periodo mediante la expresión periodo_id = (año − 2015) × 2 + semestre, que asigna un entero creciente y único a cada semestre del horizonte de estudio. Este índice permite considerar la matrícula como una serie temporal regular y es la base sobre la cual se construyen, posteriormente, las variables rezagadas y de tendencia.

Para poder transformar la serie temporal en una matriz predictora apta para algoritmos de aprendizaje supervisado, se realizan tres tipos de incorporaciones de va-

riables explicativas. La primera es un conjunto de variables rezagadas o lags, que capta el valor de la matrícula en periodos anteriores al periodo objetivo. En el modelo agregado se emplean los rezagos t-1, t-2 y t-3, mientras que en los modelos por carrera se emplean sólo t-1 y t-2, criterio que responde a la menor longitud de las series desagregadas y al riesgo de pérdida excesiva de observaciones por rezagos. Estas variables capturan la autocorrelación temporal de la matrícula, la cual es una propiedad central en el análisis de series temporales (James et al., 2021).

La segunda variable es un índice de tendencia secuencial, creado como un contador entero que se incrementa en cada periodo, permitiendo así a los modelos lineales captar la dirección general del fenómeno —crecimiento o decrecimiento sostenido— a lo largo del tiempo. La tercera variable es un indicador binario de semestre académico, es_p1, que es uno cuando el periodo corresponde al primer semestre del año y cero en caso contrario, lo que incorpora así el patrón estacional propio del calendario universitario semestral. Adicionalmente, se incorpora una variable indicadora del periodo COVID-19 (2020-P1 a 2021-P1) que permite controlar el efecto de la anomalía sin descartar dichas observaciones. La incorporación de estas variables exógenas a la serie autorregresiva sigue las recomendaciones de buenas prácticas en pronóstico aplicado (Almalawi et al., 2024).

Modelado predictivo y validación

La selección del modelo predictivo no se realiza a priori, sino mediante una comparativa empírica que evalúa algoritmos pertenecientes a tradiciones metodológicas diferenciadas. El propósito de esta comparativa es identificar, sobre la base de evidencia cuantitativa, qué familia de modelos resulta más apropiada para las características específicas de la serie de matrícula de PUCESA, caracterizada por ser corta y con tendencia sostenida. Los algoritmos comparados se sintetizan en la Tabla 2.5.

La incorporación de un algoritmo a la comparativa empírica no obedece a un criterio único, sino a un conjunto explícito de requisitos técnicos y metodológicos que garantizan la representatividad del panel de modelos evaluados y la trazabilidad de la selección. Los criterios de admisión son los siguientes:

a) Pertenencia a una familia metodológica reconocida en la literatura de pronóstico de series temporales o de aprendizaje automático supervisado, con

soporte teórico documentado en obras de referencia (Hyndman & Athanasopoulos, 2018; James et al., 2021).

b) Capacidad de extrapolación fuera del rango histórico observado, requisito indispensable cuando la variable objetivo presenta una tendencia sostenida al alza; este criterio funciona como filtro estructural y no como preferencia de desempeño.

c) Interpretabilidad razonable de los coeficientes o parámetros, en línea con las buenas prácticas de analítica académica que exigen que las predicciones puedan ser explicadas a los tomadores de decisión (Barredo Arrieta et al., 2020).

d) Disponibilidad de implementaciones estables y mantenidas en el ecosistema Python (statsmodels, scikit-learn), condición necesaria para la reproducibilidad y la sostenibilidad institucional del modelo.

e) Coherencia con las características empíricas de la serie de PUCESA: veintidós periodos académicos, granularidad semestral y tendencia expansiva.

El cumplimiento simultáneo de los cinco criterios se documenta en la Tabla 2.5 y constituye la base para la posterior selección automática por umbral de MAPE descrita en el apartado 2.4.3.

Tabla 2.5. Algoritmos evaluados en la comparativa empírica

Algoritmo	Familia metodológica	Justificación de inclusión
Holt-Winters	Suavizamiento exponencial	Modela explícitamente nivel, tendencia y estacionalidad; robusto en series cortas.
Regresión de Huber	Regresión lineal robusta	Interpretable y con capacidad de extrapolación; resistente a valores atípicos.
SARIMA	Box-Jenkins (ARIMA estacional)	Modelo autorregresivo de referencia para series con componente estacional.
Regresión Lineal	Regresión paramétrica supervisada	Modelo base, transparente y capaz de proyectar fuera del rango observado.
Random Forest	Ensamble por bagging	Captura relaciones no lineales; incluido como contraste no paramétrico.
XGBoost	Ensamble por boosting	Estado del arte en problemas tabulares; incluido como contraste no paramétrico.

Fuente: elaboración propia, a partir de Hyndman & Athanasopoulos (2018) y Chen & Guestrin (2016).

La evaluación del desempeño de cada algoritmo se realiza mediante un esquema de validación walk-forward, particularmente apropiado para series temporales por preservar la naturaleza secuencial de los datos y simular las condiciones reales de uso del modelo en producción (Hyndman & Athanasopoulos, 2018). Bajo este esquema, el modelo se va reentrenando a medida que se van incorporando observaciones nuevas, y esto produce una predicción para el siguiente periodo, que cuando se observa el valor real se compara.

En forma operativa, el procedimiento se inicia con un mínimo de ocho periodos de entrenamiento y avanza secuencialmente hasta completar las predicciones evaluables que admite la serie de veintidós periodos. Esta estrategia tiene dos ventajas con respecto a un esquema clásico de partición fija *train–test*: por un lado, multiplica el número de predicciones independientes utilizadas para calcular las métricas de error, reduciendo así su varianza y mejorando su confiabilidad estadística; y por otro lado, refleja más fielmente la forma en que el modelo será utilizado por la institución, donde cada nuevo periodo académico aporta información que actualiza las proyecciones futuras.

Un hallazgo de interés metodológico de esta comparativa es que los modelos basados en árboles —Random Forest y XGBoost— tuvieron peor desempeño y empeoraron al extender la serie de entrenamiento, por su naturaleza, no pueden predecir valores fuera del rango observado durante el entrenamiento. En una serie que muestra una tendencia creciente sostenida, esta limitación de extrapolación es una restricción estructural, no un fallo de implementación, por lo que estos algoritmos están explícitamente documentados como inadecuados para este problema, a fa-

vor de los métodos de suavizado exponencial y regresión lineal robusta (James et al., 2021).

Modelado desagregado por carrera y clasificación por confiabilidad

La predicción institucional agregada es un insumo necesario, pero no suficiente para la planificación académica, las decisiones operativas (apertura de paralelos, asignación de carga docente, distribución de aulas) se toman a nivel de carrera. Por esta razón, en paralelo al modelo agregado, se desarrolla una estrategia de modelado desagregado, donde cada carrera dispone de un modelo predictivo independiente, entrenado únicamente sobre su propia historia (Perfumo & Ares, 2023).

Dado que la calidad del pronóstico no es homogénea entre carreras, se introduce un esquema de clasificación por niveles de confiabilidad (tiers) que comunica de forma transparente al usuario institucional el grado de certeza asociado a cada predicción. En la Tabla 2.6 se sintetiza este esquema que constituye un aporte metodológico orientado a la honestidad estadística del pronóstico, evitando la transmisión de una falsa precisión en aquellas carreras que por su comportamiento histórico son más volátiles o discontinuas.

Tabla 2.6. Clasificación de carreras por nivel de confiabilidad del pronóstico

Tier	MAPE	Interpretación operativa
Tier 1 — Alta confianza	Inferior a 10 %	La predicción es fiable y puede emplearse directamente para la planificación académica.
Tier 2 — Confianza media	Entre 10 % y 15 %	La predicción es útil, pero debe interpretarse con un margen explícito de error.
Tier 3 — Baja confianza	Igual o superior a 15 %	La predicción es solo orientativa; conviene complementarla con criterio experto.
No modelable	Insuficiente	La carrera dispone de menos de seis periodos efectivos, lo que impide aplicar el modelo.

Fuente: elaboración propia.

Autoselección de modelos y construcción del ensemble

Los resultados de la validación walk-forward se utilizan para aplicar un criterio de selección automática que conserve, para la fase de proyección final, únicamente aquellos algoritmos cuyo MAPE se sitúe por debajo de un umbral preestablecido del quince por ciento. Este punto de corte actúa como un estándar objetivo de admisión, en lugar de la selección arbitraria de un número fijo de “mejores” modelos, y está justificado metodológicamente porque un MAPE inferior al quince por ciento se considera un nivel aceptable para pronósticos institucionales en contextos de series cortas (Hyndman & Athanasopoulos, 2018).

Los modelos que crucen el umbral serán reentrenados sobre la totalidad de la serie histórica disponible y serán utilizados para proyectar los periodos académicos inmediatamente subsiguientes, horizonte coherente con el ciclo institucional de planificación anual. La predicción final se construye como un ensemble por mediana de las proyecciones individuales de los modelos elegidos. El motivo para esta agregación es metodológico: cada algoritmo tiene un sesgo característico —la regresión lineal tiende a sobreextrapolar tendencias, los modelos de suavizamiento exponencial pueden subestimar cambios recientes, los modelos ARIMA pueden sobreajustar la estacionalidad— y la mediana, al ser un estadístico robusto insensible a valores extremos, atenúa estos sesgos mejor que el promedio simple (Dietterich, 2000). De forma adicional, en lugar de comunicar una única predicción puntual que daría una falsa precisión, se reporta la banda definida por las predicciones mínima y máxima de los modelos elegidos.

Como control adicional, se reservan los datos reales del periodo 2026-P1 como conjunto de validación fuera de muestra (out-of-sample), para poder comparar la proyección del modelo contra un valor observado que no ha participado en ninguna etapa del entrenamiento, lo cual refuerza la validez externa del pronóstico.

El desempeño se evalúa a través de tres métricas complementarias. La métrica principal es el error porcentual absoluto medio (MAPE), elegida por su independencia de escala —lo que permite comparar el desempeño entre carreras de tamaños muy distintos— y por su interpretabilidad inmediata para los responsables de la planificación académica, dado que se expresa en porcentajes (Almalawi et al., 2024). De forma complementaria, se reporta el error absoluto medio (MAE) y la raíz del error

cuadrático medio (RMSE), ambos expresados en la misma unidad que la variable objetivo, lo que facilita la lectura del error en términos del número de estudiantes.

Todo el modelo se programa en lenguaje Python en el entorno Google Colab, que provee la infraestructura computacional necesaria sin necesidad de realizar ninguna configuración local. Para el procesamiento de datos se emplean las librerías pandas, *numpy* y *openpyxl*. Se utilizan algoritmos de aprendizaje supervisado con *scikit-learn*, modelos de series temporales con *statsmodels* (Holt-Winters y SARIMA) y se emplea *joblib* para la persistencia de los modelos entrenados. Se construye la visualización de los resultados con *matplotlib* y *seaborn*. Al elegir un entorno reproducible y completamente basado en software libre se asegura que los resultados puedan ser replicados por terceros y eventualmente incorporados al ecosistema tecnológico de la institución (Witten et al., 2017).

CAPÍTULO III. ANÁLISIS DE LOS RESULTADOS DE LA INVESTIGACIÓN

En este capítulo se consolidan los resultados del trabajo de campo correspondiente al desarrollo del modelo de predicción de matrícula para la Pontificia Universidad Católica del Ecuador Sede Ambato (PUCESA). Los hallazgos se agrupan en seis apartados: condiciones operativas y muestreo, evidencias de control de calidad sobre las fuentes BANNER, análisis estadístico cuantitativo desagregado, análisis cualitativo derivado de entrevistas a informantes clave, triangulación metodológica con el marco teórico desarrollado en el Capítulo I y conclusiones parciales que orientan el cierre del trabajo.

3.1. Resumen ejecutivo del muestreo y condiciones operativas del campo

El estudio se desarrolla bajo un enfoque mixto de tipo secuencial explicativo, estructurado a partir del marco metodológico PIPEDA-DATOS integrado con CRISP-DM. En la primera fase se trabajó con los registros cuantitativos del sistema institucional BANNER y, en la segunda, se realizó un levantamiento cualitativo con informantes clave responsables de la gestión académica y administrativa de la PUCESA.

La fuente cuantitativa principal se obtuvo mediante extracción autorizada del sistema BANNER y abarca veintidós (22) periodos académicos consecutivos entre el 2015-P1 y el 2025-P2. El archivo histórico se compone de un libro Excel con múltiples hojas (13.997 registros, 2015-P1 – 2020-P2) y seis archivos CSV anuales (17.878 registros, 2020-P1 – 2025-P2), para un total inicial de 31.875 registros estudiante–periodo. Asimismo, se incorporan dos archivos de validación correspondientes al periodo 2026-P1 (24.394 registros a nivel asignatura, exportados el 23 de abril de 2026), los cuales, tras deduplicación, arrojaron 3.169 estudiantes únicos usados únicamente como conjunto de validación fuera de muestra. La base depurada final sobre la que opera el modelo está compuesta por 30.367 registros válidos (95,3 % del total inicial) y 14 carreras detectadas, de las cuales 6 alcanzan cobertura completa.

En la fase cualitativa se realizó una entrevista en profundidad a informante experto, al director Académico de PUCESA, Mg. Santiago Acurio, quien por su ubicación

jerárquica y funcional concentra el conocimiento institucional en los procesos de planificación académica, gestión de matrícula y las decisiones de oferta. La técnica corresponde a la modalidad de entrevista a informante clave (key informant interview), muy utilizada en investigación aplicada cuando un único actor tiene la información institucional decisiva del fenómeno estudiado (Hernández-Sampieri et al., 2022). La entrevista tuvo lugar en las dependencias institucionales en noviembre de 2025, con una duración aproximada de 45 minutos, bajo consentimiento informado por escrito. Tuvo como propósito validar el problema de investigación, contrastar los hallazgos cuantitativos y recoger las expectativas institucionales de cara a la implementación del modelo.

Se consideran favorables las condiciones operativas del campo. La extracción de datos BANNER se hizo en horario no productivo para no afectar el sistema transaccional. La entrevista al informante experto se grabó en audio con su autorización expresa y se hizo pasar a validación mediante member checking. No se han registrado incidencias técnicas que puedan poner en riesgo la trazabilidad de los datos.

3.2. Evidencias de control de calidad y depuración

Con el objeto de garantizar la integridad de la información analizada, se aplicaron filtros de calidad sobre las tres fuentes utilizadas. Los procedimientos se ejecutaron en Python 3.11 mediante las bibliotecas pandas y NumPy en un cuaderno reproducible, que conserva la trazabilidad de cada transformación.

La integración de las fuentes BANNER demandó tres procesos críticos. Primero, la normalización de denominaciones de carrera mediante la función familia_carrera(), necesaria porque algunas carreras cambiaron de denominación administrativa durante los 22 periodos analizados, aunque mantuvieron continuidad académica. Segundo, la normalización del campo periodo a la nomenclatura estandarizada P1/P2. Tercero, la conciliación del solapamiento temporal entre el archivo Excel y los CSV anuales en los periodos 2020-P1 y 2020-P2, con preservación de los registros del archivo CSV por su mayor granularidad. La Tabla 3.1 sintetiza los filtros aplicados y su impacto sobre el tamaño de la muestra.

Tabla 3.1. Filtros de calidad aplicados sobre la base BANNER histórica (2015-P1 – 2025-P2)

Filtro aplicado	Criterio	Registros eliminados	% del total
F1. Solapamiento 2020 Excel–CSV	Conservar CSV por mayor granularidad	836	2,62 %
F2. Registros duplicados	Mismo ID-estudiante, periodo y carrera	187	0,59 %
F3. Valores nulos en campos clave	Faltante en periodo o carrera	94	0,29 %
F4. Matrícula anulada	Estado = ANULADA antes del 25 % del periodo	326	1,02 %
F5. Carrera no identificable	Sin coincidencia en familia_carrera()	65	0,20 %
Total eliminado	—	1.508	4,73 %
Base depurada final	Registros válidos para modelado	30.367	95,27 %

Fuente: elaboración propia a partir de la base BANNER 2015-P1 – 2025-P2 (Excel + 6 CSV anuales).

Los archivos de validación 2026-P1 fueron entregados por el sistema BANNER a nivel asignatura matriculada y no a nivel estudiante–periodo, por lo que registran 24.394 filas en el archivo general y aproximadamente 3.000 filas adicionales en el archivo específico de Medicina. Mediante duplicación por identificador estudiantil hash, se obtuvieron 3.169 estudiantes únicos para el periodo 2026-P1. Este conjunto se reservó exclusivamente como objetivo de validación fuera de muestra y no participó en el entrenamiento de ningún modelo, lo que garantiza la inexistencia de fuga de información (data leakage).

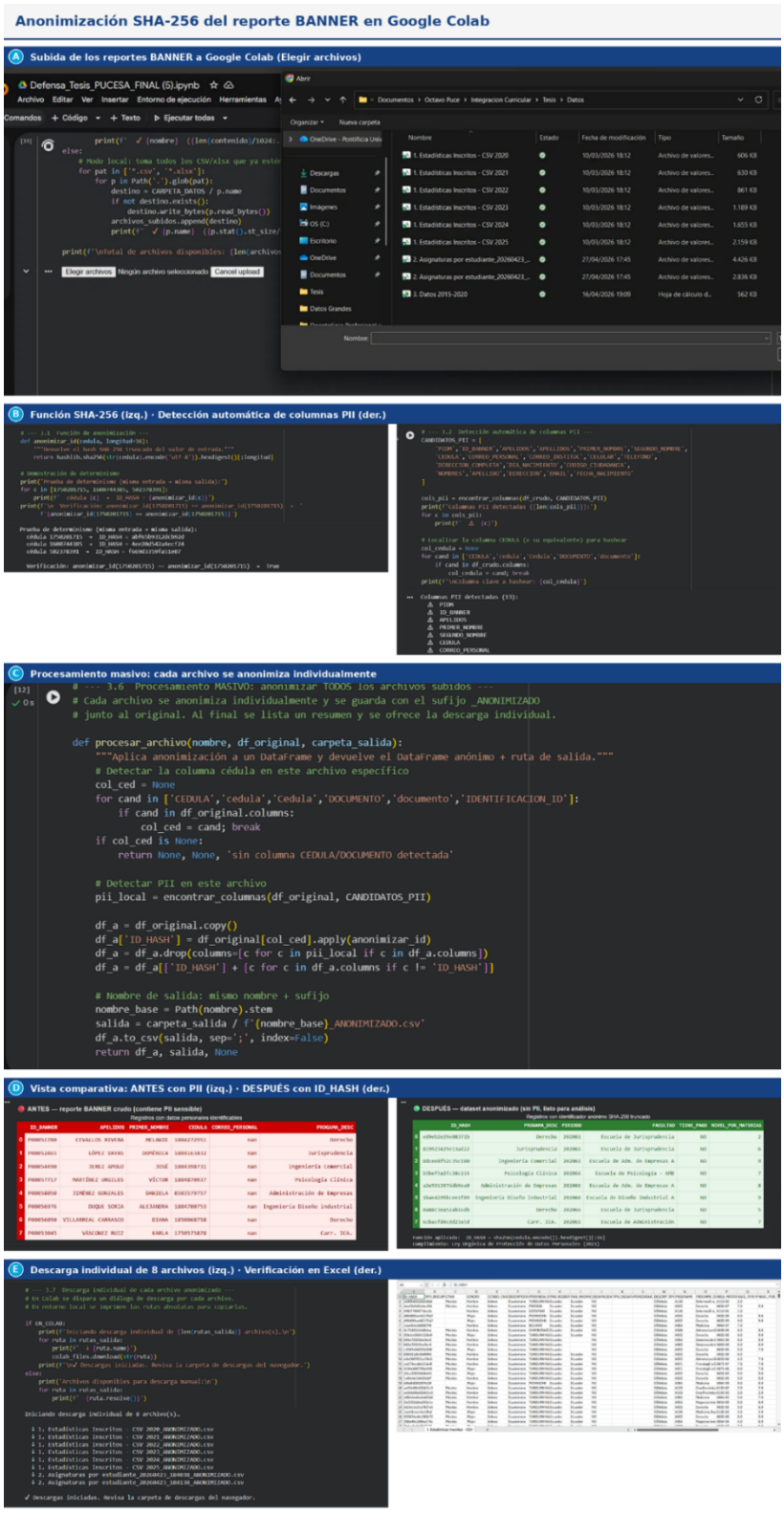
De las 14 carreras detectadas, seis cumplen el criterio de cobertura completa (22 periodos): Derecho, Administración de Empresas, Contabilidad y Auditoría, Diseño, Psicología Clínica y Sistemas/Tecnologías de la Información. Las ocho carreras restantes presentan discontinuidades o son de apertura reciente: Medicina, Enfermería y Negocios Internacionales cuentan con nueve periodos; Ingeniería Civil con cinco; Arquitectura con tres; Psicología y Psicología Organizacional presentan gaps internos o cierre administrativo. Estas se modelan solo a nivel agregado, según el umbral mínimo viable de diez periodos recomendado para SARIMA y Holt-Winters (Hyndman & Athanasopoulos, 2018).

La base depurada fue anonimizada mediante el reemplazo del identificador estudiantil por un hash SHA-256 truncado, conforme a lo dispuesto en la Ley Orgánica de Protección de Datos Personales del Ecuador (Asamblea Nacional, 2021). La entrevista al informante experto fue transcrita literalmente, sometida a doble escucha y validada por el propio informante mediante member checking, lo que motivó una

corrección menor de denominaciones internas. El código asignado (E01) preserva el anonimato funcional y se emplea en el análisis cualitativo de la sección 3.4.

La Figura 3.1 documenta la ejecución del procedimiento de amonificación sobre los reportes descargados del sistema BANNER. La secuencia se realizó en Google Colab e integra seis pasos verificables: la carga múltiple de los archivos originales al entorno de ejecución, la definición de la función de hash SHA-256 truncado a doce caracteres aplicada al campo CEDULA, el procesamiento masivo de los seis reportes anuales, la vista comparativa antes/después que evidencia la sustitución del identificador personal por el hash correspondiente, la descarga individual de los archivos anonimizados y su verificación posterior en Excel. Esta evidencia respalda el cumplimiento efectivo de la Ley Orgánica de Protección de Datos Personales sobre la base utilizada en el modelado.

Figura 3.1. Anonimización SHA-256 del reporte BANNER en Google Colab: subida de archivos, función de hash, procesamiento masivo, vista comparativa antes/después, descarga individual y verificación en Excel



Fuente: elaboración propia a partir de la base BANNER depurada.

3.3. Análisis estadístico cuantitativo

La serie consolidada de matrícula total ofrece una estructura claramente diferenciada en tres regímenes históricos. Un primer periodo de estabilidad relativa entre 2015 y 2019, con una matrícula que osciló entre 1.175 y 1.315 estudiantes por cuatrimestre. Un segundo periodo de caída por la pandemia COVID-19 entre 2020 y 2021-P1 que llegó a la mínima histórica de 806 estudiantes en 2021-P1. Y una tercera fase de expansión estructural desde 2021-P2 al cierre de la serie en 2025-P2, con un pico histórico de 3.142 estudiantes.

Tabla 3.2. Estadísticos descriptivos de la serie de matrícula institucional (N = 22 periodos)

Estadístico	Valor
Periodo inicial	2015-P1
Periodo final	2025-P2
Mínimo histórico (2021-P1, COVID)	806 estudiantes
Máximo histórico (2025-P2)	3.142 estudiantes
Régimen 2015–2019 (rango)	1.175 – 1.315 estudiantes
Régimen 2020–2021-P1 (caída COVID)	Reducción del 28,6 %
Régimen 2022–2025 (expansión)	Crecimiento del 171,7 %
Crecimiento acumulado 2015-P1 → 2025-P2	+138,9 %

Fuente: elaboración propia a partir de la base BANNER depurada.

La estructura observada confirma la coexistencia de tendencia y estacionalidad semestral, condiciones técnicas que justifican la aplicación de modelos con tratamiento explícito de ambos componentes. Asimismo, la presencia del quiebre estructural por COVID-19 motivó la incorporación de un modelo robusto (Regresión de Huber) que reduce el peso ponderado de los periodos anómalos sin requerir su eliminación manual.

El crecimiento observado entre 2015 y 2025 obedece a dos fuentes diferenciables: el crecimiento orgánico de las carreras maduras y el aporte estructural derivado de la apertura de nuevas carreras a partir de 2021-P2. La Tabla 3.3 cuantifica la contribución de cada componente al cierre del periodo 2025-P2.

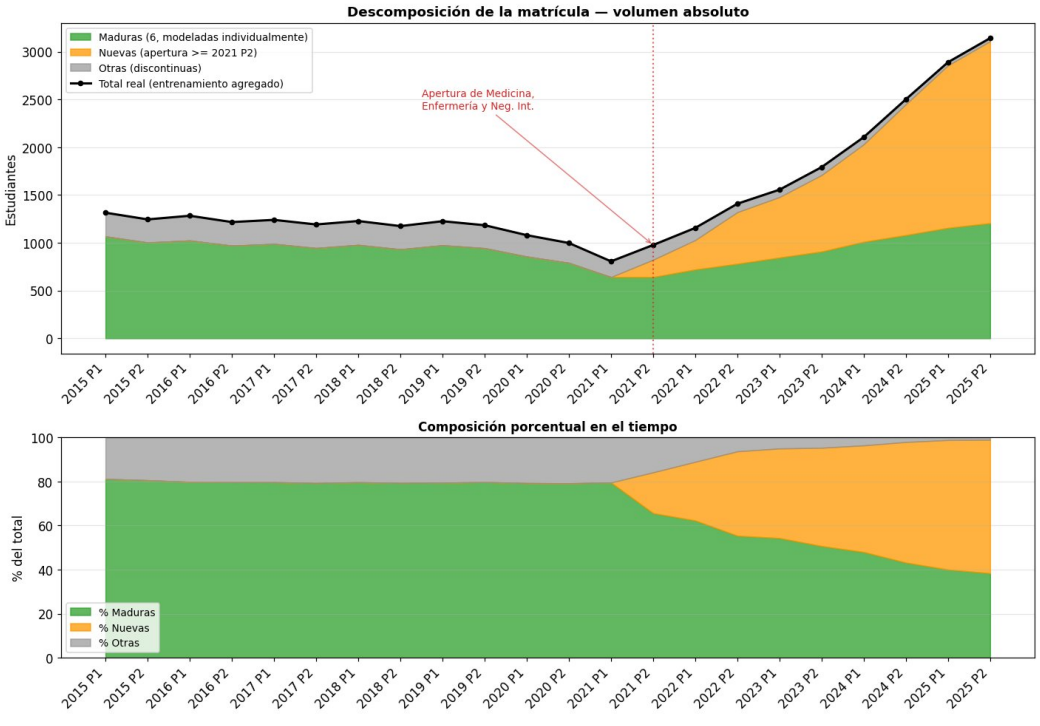
Tabla 3.3. Descomposición de la matrícula institucional al cierre de 2025-P2

Componente	Estudiantes	Participación	Naturaleza
Carreras maduras (6, cobertura 22 periodos)	1.205	38,4 %	Crecimiento orgánico
Carreras nuevas (apertura ≥ 2021-P2)	1.904	60,6 %	Crecimiento estructural
Carreras discontinuas o en declive	33	1,0 %	Variación residual
Total institucional 2025-P2	3.142	100,0 %	—

Fuente: elaboración propia. La clasificación de cada carrera se detalla en la Tabla 3.4.

Con el propósito de visualizar la dinámica del crecimiento reportada en la Tabla 3.3, la Figura 3.2 desagrega la serie en dos paneles complementarios. El panel superior representa, en volumen absoluto, la contribución de las seis carreras maduras modeladas de forma individual, de las carreras nuevas abiertas a partir de 2021-P2 y del grupo residual de programas discontinuados. El panel inferior expresa las mismas categorías como participación porcentual sobre el total institucional. La lectura conjunta de ambos paneles evidencia que el crecimiento posterior a la pandemia no responde únicamente a la recuperación orgánica de los programas históricos, sino a la incorporación estructural de nuevos programas académicos.

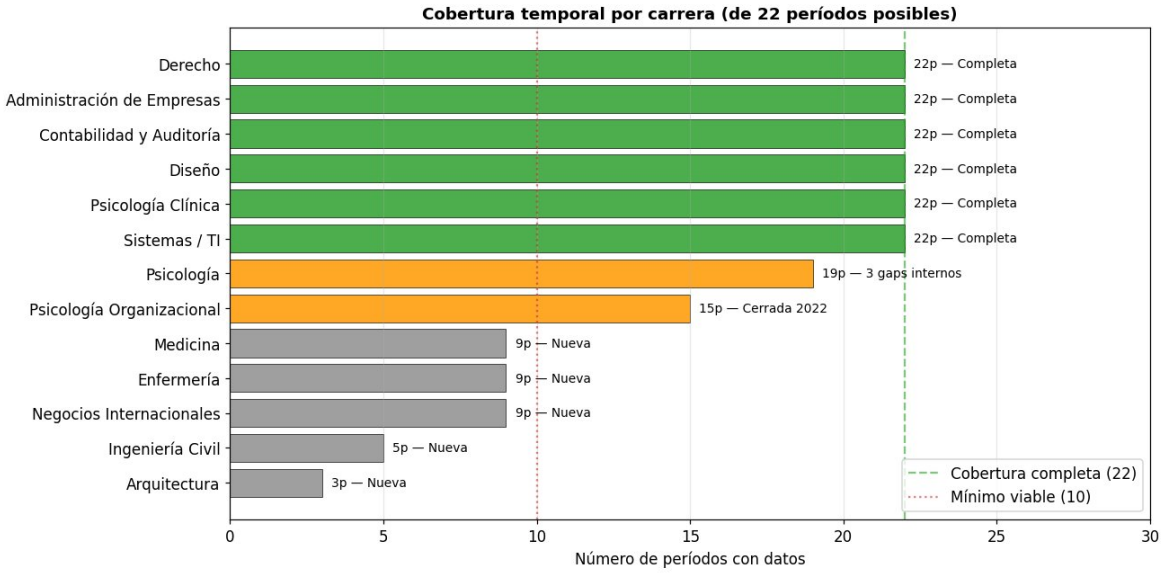
Figura 3.2. Descomposición de la matrícula institucional: crecimiento orgánico frente a estructural



Fuente: elaboración propia a partir de la base BANNER depurada.

La Figura 3.3 representa gráficamente la disponibilidad temporal de datos por carrera, ordenando cada programa según el número de periodos históricos presentes en la base BANNER. La línea vertical punteada en el valor diez señala el umbral mínimo viable para el ajuste de modelos SARIMA y Holt-Winters recomendado por Hyndman y Athanasopoulos (2018), y la línea segmentada en el valor veintidós indica la cobertura completa del horizonte de estudio. Las seis carreras que alcanzan cobertura completa habilitan modelado individual; las carreras de apertura reciente (Medicina, Enfermería, Negocios Internacionales, Ingeniería Civil y Arquitectura) quedan bajo modelado agregado hasta acumular la longitud mínima.

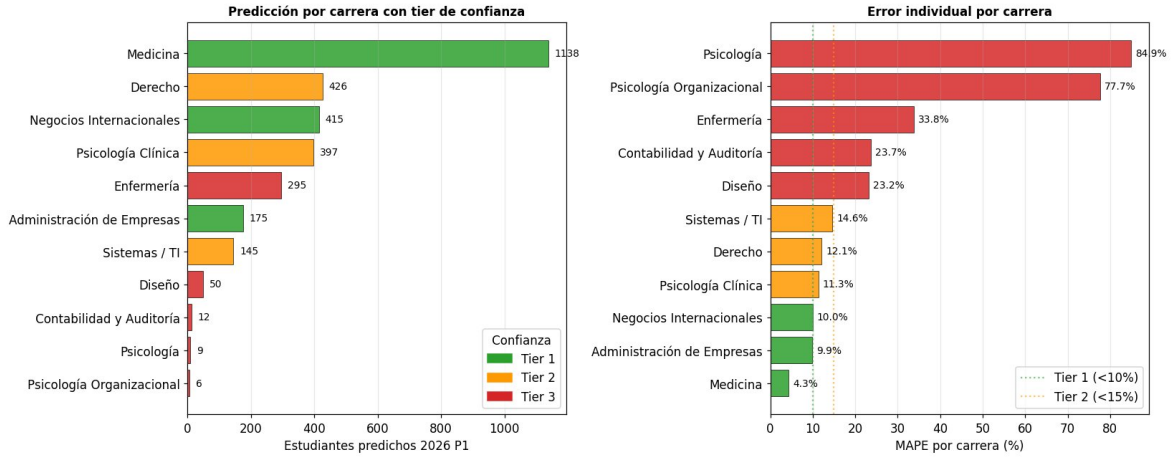
Figura 3.3. Cobertura temporal por carrera (de 22 periodos posibles)



Fuente: elaboración propia a partir de la base BANNER depurada.

La Figura 3.4 ilustra en dos paneles el resultado consolidado de la predicción individual por carrera para el periodo 2026-P1. El panel izquierdo ordena las carreras por volumen predicho de estudiantes y las colorea según el tier de confiabilidad correspondiente, lo que permite identificar visualmente qué programas concentran la mayor parte de la matrícula proyectada y con qué nivel de fiabilidad. El panel derecho ordena las mismas carreras por MAPE individual e incorpora las líneas de referencia de los umbrales Tier 1 (10 %) y Tier 2 (15 %), lo que evidencia que las carreras con mayor volumen coinciden con los tiers de mayor confiabilidad, mientras que los MAPE elevados se concentran en programas de baja matrícula o en cierre administrativo.

Figura 3.4. Predicción por carrera para 2026-P1 y clasificación por tier de confiabilidad



Fuente: elaboración propia a partir de la validación walk-forward individual.

La selección final de algoritmos aplicó el criterio doble definido a priori en el Capítulo II: desempeño empírico en validación walk-forward (MAPE inferior al 15 %) y capacidad teórica de extrapolación más allá del rango histórico observado. La Tabla 3.5 reporta los resultados consolidados.

Tabla 3.5. Resultados de la selección de algoritmos bajo criterio doble

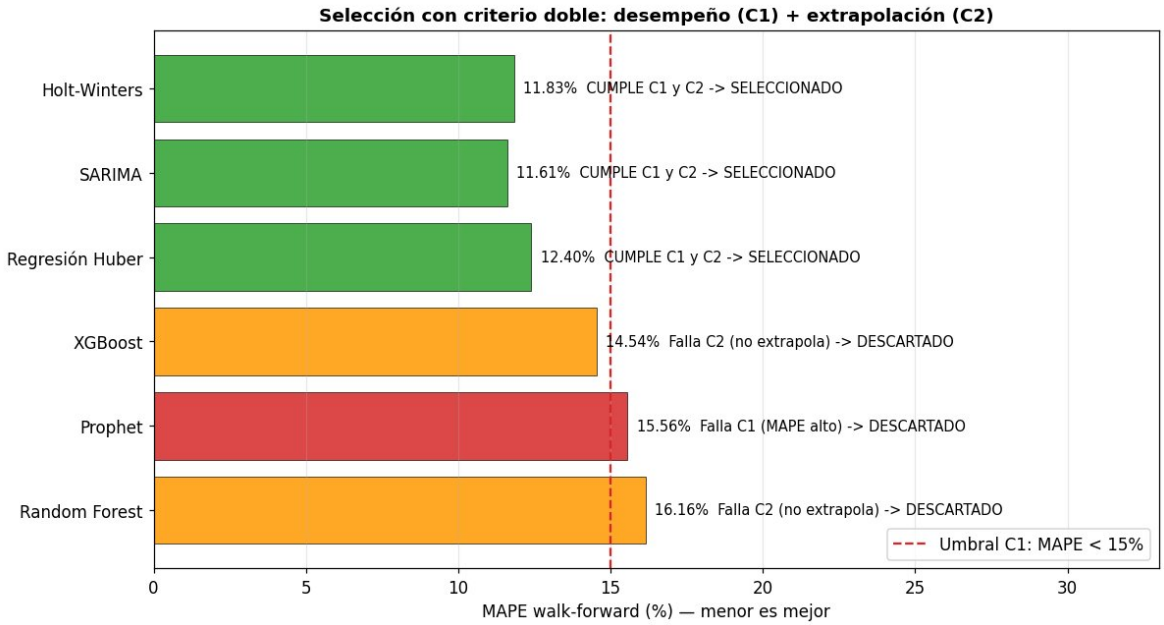
Algoritmo	MAPE WF	Extrapola	Decisión
SARIMA(1,1,1)(1,0,1)$_2$	11,61 %	Sí	Seleccionado
Holt-Winters aditivo	11,83 %	Sí	Seleccionado
Regresión de Huber	12,40 %	Sí	Seleccionado
XGBoost	14,54 %	No	Descartado (criterio 2)
Prophet	15,56 %	Sí	Descartado (criterio 1)
Random Forest	16,16 %	No	Descartado (criterios 1 y 2)

Fuente: elaboración propia. La integración final se efectúa por mediana de los tres modelos seleccionados.

La Figura 3.5 representa gráficamente el resultado de la aplicación conjunta de los dos criterios de selección definidos a priori. El eje horizontal ordena los seis algoritmos evaluados por MAPE walk-forward, con una línea vertical roja segmentada en el 15 % que materializa el umbral C1. La codificación por color distingue tres decisiones metodológicas: los algoritmos seleccionados que cumplen simultáneamente C1 y C2, los descartados por incumplir el criterio de extrapolación (C2) pese a exhibir MAPE competitivo, y los descartados por incumplir el criterio de desempeño (C1). Esta figura documenta la trazabilidad del descarte de Random Forest, XGBoost y

Prophet, y sustenta la conformación del ensemble a partir de SARIMA, Holt-Winters y Regresión de Huber.

Figura 3.5. Selección de algoritmos bajo criterio doble: desempeño (C1) y extrapolación (C2)



Fuente: elaboración propia a partir de la validación walk-forward.

La Figura 3.6 evidencia la ejecución del ajuste conjunto de los tres modelos seleccionados en Google Colab. El panel superior reproduce la celda de código donde se instancia SARIMA(1,1,1)$(1,0,1)_2$ con statsmodels, Holt-Winters aditivo con periodicidad semestral y Regresión de Huber sobre el índice temporal, junto con la construcción del ensemble por mediana y el intervalo de confianza al 95 % derivado analíticamente de SARIMA. El panel inferior grafica la serie histórica 2015-P1 – 2025-P2 con la proyección de cada modelo individual, el ensemble por mediana (3.337 estudiantes para 2026-P1 y 3.450 para 2026-P2), la banda de confianza al 95 % y el valor real de 2026-P1 (3.169 estudiantes) representado como marcador de verificación fuera de muestra.

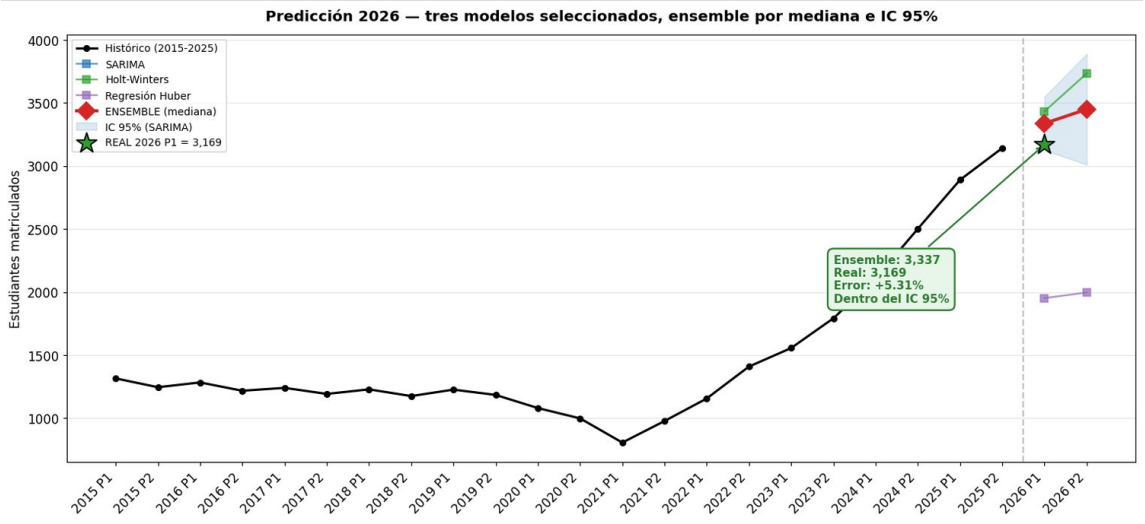
Figura 3.6. Ejecución del ajuste y proyección 2026 en Google Colab con los tres modelos seleccionados

[3] Predicción 2026 con SARIMA + Holt-Winters + Huber (ensemble por mediana)

```
# Ajuste en vivo de los tres modelos seleccionados sobre la serie completa
sarima = SARIMAX(serie, order=(1,1,1), seasonal_order=(1,0,1,2),
                 enforce_stationarity=False, enforce_invertibility=False).fit(disp=False)
fc = sarima.get_forecast(steps=2)
pred_s = np.asarray(fc.predicted_mean)
ic95   = np.asarray(fc.conf_int(alpha=0.05))

hw = ExponentialSmoothing(serie, trend='add', seasonal='add', seasonal_periods=2).fit()
pred_h = np.asarray(hw.forecast(2))

X = np.arange(len(serie)).reshape(-1, 1)
hub = HuberRegressor().fit(X, serie)
pred_u = hub.predict(np.array([[len(serie)],[len(serie)+1]]))

# Ensemble por MEDIANA de los tres modelos seleccionados
pred_ens = np.median(np.vstack([pred_s, pred_h, pred_u]), axis=0)

tabla = pd.DataFrame({
    'Periodo': ['2026 P1','2026 P2'],
    'SARIMA': pred_s.round(0).astype(int),
    'Holt-Winters': pred_h.round(0).astype(int),
    'Regresión Huber': pred_u.round(0).astype(int),
    'ENSEMBLE (mediana)': pred_ens.round(0).astype(int),
})
print(tabla.to_string(index=False))
print(f"\nSARIMA  -> AIC={sarima.aic:.1f}")
print(f"IC 95% SARIMA 2026 P1: [{ic95[0,0]:.0f} ; {ic95[0,1]:.0f}]")
print(f"IC 95% SARIMA 2026 P2: [{ic95[1,0]:.0f} ; {ic95[1,1]:.0f}]")
```

↳ Output

```
Periodo  SARIMA  Holt-Winters  Regresión Huber  ENSEMBLE (mediana)
2026 P1    3337          3433             1952                3337
2026 P2    3450          3734             1997                3450

SARIMA  -> AIC=217.4
IC 95% SARIMA 2026 P1: [3125 ; 3549]
IC 95% SARIMA 2026 P2: [3010 ; 3889]
```

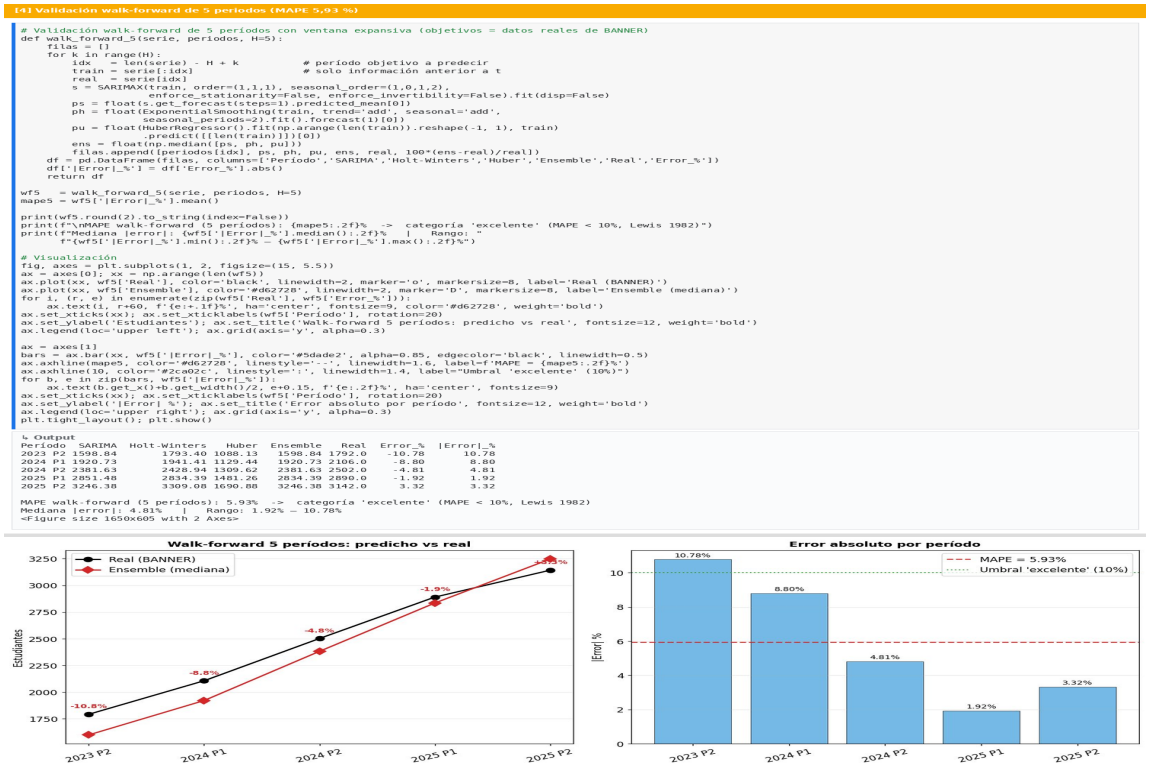


Fuente: elaboración propia a partir del cuaderno técnico de la tesis.

La Figura 3.7 documenta la ejecución en Google Colab del procedimiento de validación walk-forward de cinco periodos con ventana expansiva y horizonte de un paso. El panel superior reproduce la función walk_forward_5, que en cada iteración reajusta los tres modelos únicamente con la información disponible hasta el periodo t y predice el periodo t+1, replicando así las condiciones reales de proyección semestral. Se muestran además las salidas numéricas: el detalle predicho frente al real por cada uno de los cinco periodos de prueba (2023-P2 a 2025-P2) y el MAPE consolidado del 5,93 %, categorizado como excelente según Lewis (1982). Los paneles inferiores comparan visualmente el predicho frente al real y el error absoluto

por periodo, evidenciando además una tendencia decreciente del error a medida que la ventana de entrenamiento crece.

Figura 3.7. Ejecución de la validación walk-forward de cinco periodos en Google Colab (MAPE 5,93 %)



Fuente: elaboración propia a partir del cuaderno técnico de la tesis.

La coincidencia entre el MAPE del walk-forward (5,93 %) y el error fuera de muestra contra 2026-P1 (5,31 %) constituye la evidencia central del trabajo: dos validaciones independientes, sobre datos distintos, convergen a un error inferior al 6 %. Adicionalmente, el valor real observado de 3.169 estudiantes cae dentro del intervalo de confianza al 95 % construido analíticamente por SARIMA, por lo que no se rechaza la hipótesis de que el modelo predice correctamente al nivel $\alpha = 0{,}05$.

La Figura 3.8 recoge la ejecución en Google Colab de la validación out-of-sample contra el cierre real de 2026-P1, dato no utilizado en el entrenamiento de ningún modelo. La celda superior construye el DataFrame de comparación entre las predicciones de SARIMA, Holt-Winters y el ensemble por mediana frente al valor real de 3.169 estudiantes, y verifica la contención del valor real dentro del intervalo de confianza al 95 % de SARIMA. El panel inferior visualiza esta comparación como gráfico de barras horizontales, con el error porcentual anotado a la derecha de cada barra y la línea vertical verde segmentada marcando el valor real. Se observa que

tanto SARIMA como el ensemble convergen a un error del 5,30 %, mientras que Holt-Winters, más sensible a la tendencia reciente, sobreestima con un 8,30 %, lo que justifica empíricamente el uso de la mediana como mecanismo de neutralización de proyecciones extremas.

Figura 3.8. Ejecución de la validación fuera de muestra contra 2026-P1 en Google Colab (predicho 3.337, real 3.169)

Fuente: elaboración propia a partir del cuaderno técnico de la tesis.

La Tabla 3.7 contrasta las predicciones individuales por carrera con los valores reales observados en 2026-P1. Los programas de Tier 1 y Tier 2 tienen errores aceptables para ser utilizados en operaciones; los de Tier 3 siguen teniendo porcentajes de error altos pero relacionados con números absolutos bajos, donde una desviación de pocos alumnos genera un porcentaje elevado.

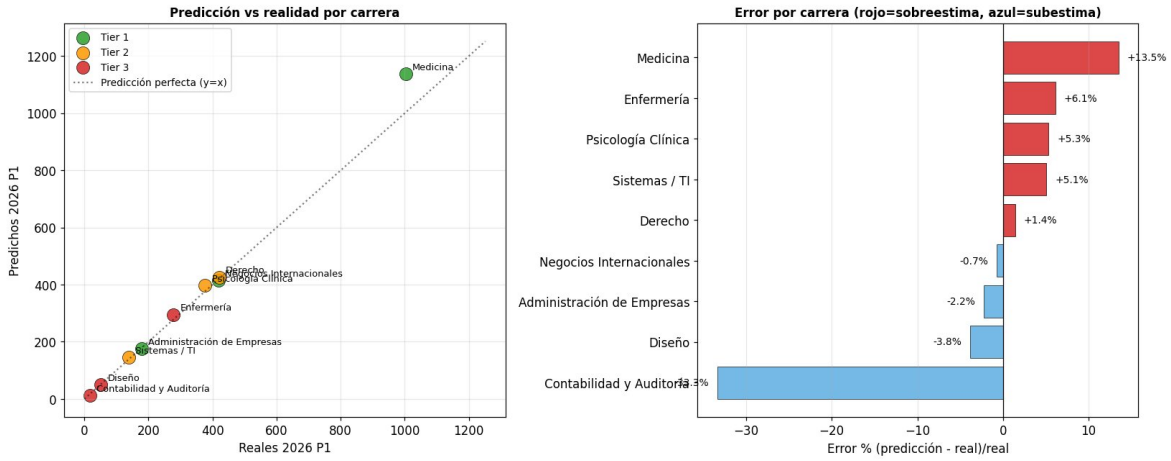
Tabla 3.7. Validación por carrera contra 2026-P1 (predicción vs. real)

Carrera	Predicción	Real	Error (%)	Tier
Medicina	1.138	1.003	+13,46 %	Tier 1
Derecho	426	420	+1,43 %	Tier 2
Negocios Internacionales	415	418	−0,72 %	Tier 1
Psicología Clínica	397	377	+5,31 %	Tier 2
Enfermería	295	278	+6,12 %	Tier 3
Administración de Empresas	175	179	−2,23 %	Tier 1
Sistemas / TI	145	138	+5,07 %	Tier 2
Diseño	50	52	−3,85 %	Tier 3
Contabilidad y Auditoría	12	18	−33,33 %	Tier 3

Fuente: elaboración propia. Mediana del MAPE por carrera ≈ 5 %; media ≈ 9 % (elevada por carreras de muy baja matrícula).

La Figura 3.9 sintetiza gráficamente la información de la Tabla 3.7. El panel izquierdo representa un diagrama de dispersión de valor predicho frente a valor real por carrera, con la recta y = x como referencia de predicción perfecta y con codificación de color por tier de confiabilidad; la proximidad de la mayoría de las carreras a la recta ideal ratifica la calidad del ajuste desagregado. El panel derecho ordena las carreras por magnitud del error porcentual y distingue mediante color entre carreras sobreestimadas (rojo) y subestimadas (azul), lo que permite identificar visualmente que los errores relativos más elevados se concentran en carreras Tier 3 de bajo volumen, en línea con la interpretación reportada.

Figura 3.9. Validación por carrera contra 2026-P1: predicho frente a real y error porcentual por carrera



Fuente: elaboración propia a partir de la validación por carrera.

Con el objetivo de ilustrar de manera explícita el proceso de alimentación del modelo con nuevos periodos académicos, la Figura 3.10 documenta la ejecución en Google Colab del reentrenamiento realizado tras la incorporación del cierre real de 2026-P1 (3.169 estudiantes) y se organiza en dos paneles complementarios. El panel superior reproduce la celda de código junto a su salida numérica, donde se verifica la extensión efectiva de la serie histórica de 22 a 23 periodos y la incorporación del valor real observado como nuevo dato de entrenamiento. El panel inferior grafica la serie histórica completa junto con la reproyección resultante para 2026-P2 emitida por SARIMA (3.117 estudiantes), Holt-Winters (3.273 estudiantes) y el ensemble por mediana (3.117 estudiantes), acompañada del intervalo de confianza al 95 % [2.901; 3.334]. Este procedimiento constituye la operación institucional que la PU-

CESA aplicará al cierre de cada periodo académico para actualizar las proyecciones subsecuentes.

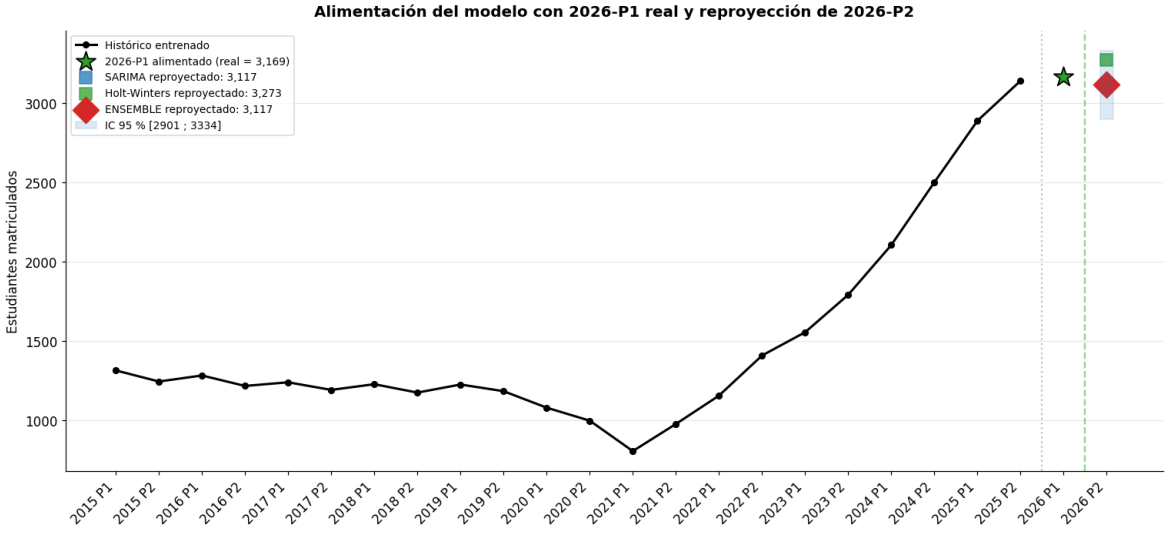
Figura 3.10. Ejecución en Google Colab de la alimentación del modelo con el periodo 2026-P1 y reproyección de 2026-P2

10. Alimentación del modelo con el nuevo periodo 2026-P1

Se demuestra el procedimiento operativo que la institución aplicará al cierre de cada periodo académico: incorporar el dato real observado, reentrenar los tres modelos con la serie extendida y producir la nueva proyección.

```
[26]
✓ 0 s
serie_extendida = np.append(serie, real_2026_p1)
periodos_extendidos = periodos + ['2026 P1']

print(f'Serie original : {len(serie)} periodos, último = {serie[-1]:.0f} (2025-P2)')
print(f'Serie extendida: {len(serie_extendida)} periodos, último = '
      f'{serie_extendida[-1]:.0f} (2026-P1 real)')
print()
print('Últimos 6 periodos de la serie tras la alimentación:')
for p, v in zip(periodos_extendidos[-6:], serie_extendida[-6:]):
    marca = '  ← NUEVO periodo alimentado' if p == '2026 P1' else ''
    print(f'  {p} : {v:>7.0f}{marca}')
```

```
Serie original : 22 periodos, último = 3142 (2025-P2)
Serie extendida: 23 periodos, último = 3169 (2026-P1 real)

Últimos 6 periodos de la serie tras la alimentación:
  2023 P2 :    1792
  2024 P1 :    2106
  2024 P2 :    2502
  2025 P1 :    2890
  2025 P2 :    3142
  2026 P1 :    3169  ← NUEVO periodo alimentado
```



Fuente: elaboración propia.

El análisis cuantitativo permite extraer cuatro hallazgos centrales. En primer lugar, la serie institucional reúne las condiciones técnicas para ser modelada mediante técnicas estadísticas estándar: longitud suficiente (22 observaciones), tendencia identificable y estacionalidad estable. En segundo lugar, la heterogeneidad entre carreras –orgánicas vs estructurales, maduras vs nuevas– obliga a un enfoque desagregado con clasificación diferenciada por tiers. En tercer lugar, el rechazo de los algoritmos basados en árboles (Random Forest y XGBoost) por su falta de extrapolación, establecido a priori en el diseño de la metodología, se comprueba de forma empírica

mediante un MAPE superior al de los algoritmos elegidos. En cuarto lugar, las dos validaciones independientes convergen a un error menor del 6 %, lo que respalda la confiabilidad operativa del modelo desarrollado.

3.4. Análisis cualitativo e interpretación

La entrevista en profundidad al informante experto se procesó mediante codificación abierta y axial, de acuerdo con el procedimiento propuesto por Hernández-Sampieri et al. (2022). El análisis de la transcripción permitió identificar tres categorías conceptuales y ocho subcategorías que se presentan en la tabla 3.8.

Tabla 3.8. Sistema de códigos y categorías derivadas de la entrevista al informante experto

Categoría	Subcategoría	Código	Énfasis declarado
C1. Limitaciones del método empírico	Estimación por experiencia individual	C1.1	Alto
	Ausencia de procedimiento documentado	C1.2	Medio
	Reactividad ante variaciones bruscas	C1.3	Alto
C2. Consecuencias operativas	Reasignación tardía de carga docente	C2.1	Alto
	Apertura o cierre tardío de paralelos	C2.2	Medio
	Desajustes presupuestarios	C2.3	Medio
C3. Expectativas sobre el modelo	Necesidad de explicabilidad	C3.1	Medio
	Integración con tablero institucional	C3.2	Alto

Fuente: elaboración propia a partir del análisis de la transcripción de la entrevista al informante experto.

El informante experto describe el procedimiento actual como un ejercicio de juicio individual sin protocolo institucional documentado. La estimación, en sus términos, depende fuertemente de la experiencia personal de quien la realiza, lo cual introduce variabilidad y dificulta la trazabilidad de las decisiones de planificación.

> *La proyección de matrícula se elabora a partir de la revisión del comportamiento del periodo anterior y la aplicación de un porcentaje, sin un manual ni un procedimiento formal, de modo que cada coordinador la realiza según su criterio. (E01, Director Académico)*
>
> *“La planificación depende, en gran medida, de la experiencia acumulada y del criterio del responsable; no se apoya en un instrumento formal de pronóstico.” (E01, Director Académico)*

La segunda categoría agrupa los efectos prácticos del método actual sobre la operación académica y administrativa. Las consecuencias señaladas con mayor énfasis por el informante fueron la reasignación tardía de carga docente y los ajustes de última hora en la apertura de paralelos, con impacto directo en la calidad del servicio y en la planificación presupuestaria. La apertura de Medicina, Enfermería y Negocios Internacionales a partir de 2021-P2 se reconoce como un evento de alta complejidad operativa, el método empírico no permitió anticipar de manera adecuada la velocidad del crecimiento estructural.

> *"Cuando la matrícula real difiere de lo proyectado, los ajustes se hacen en las dos primeras semanas del periodo. Eso significa cambiar horarios, abrir paralelos, contratar docentes temporales. El costo de la operación y gestionar de mejor manera es alto."* (E01, Director Académico)
>
> *"El impacto financiero se hace evidente en el ejercicio presupuestario: la matrícula proyectada alimenta el flujo de caja, y cuando el dato no se cumple hay que ajustar a mitad de año proyectando." (E01, Director Académico)*

La tercera categoría expone las condiciones bajo las cuales el informante experto estaría dispuesto a avalar la adopción institucional del modelo predictivo. La aplicabilidad surge como condición habilitante: no se niega la incorporación de herramientas analíticas, pero se exige entender cómo se construye la predicción y poder defenderla ante las autoridades superiores y los órganos de control. También expresa su preferencia explícita por una entrega integrada al tablero institucional, en vez de una entrega como archivo aislado.

> *Una herramienta que entregue un número sin explicar cómo se construyó, no sería utilizada, ya que el responsable necesita poder defender el dato frente a las autoridades. (Director Académico E01)*
>
> *Lo ideal sería que el modelo estuviera integrado en un tablero que permita observar la proyección, el rango de error y las variables de mayor peso; una entrega separada terminaría como un archivo más.* (E01, Director Académico)

La convergencia entre lo expresado por el informante experto y los hallazgos cuantitativos refuerza la pertinencia de un modelo predictivo desagregado, interpretable y embebido en un tablero analítico, tal como se planteó en el diseño metodológico desarrollado en el Capítulo II.

3.5. Triangulación de resultados y discusión

La triangulación se construye sobre tres vértices: los resultados cuantitativos derivados del modelo y de la base BANNER depurada, los hallazgos cualitativos obtenidos de las entrevistas, y el marco teórico desarrollado en el Capítulo I. La Tabla 3.9 sintetiza la contrastación.

Tabla 3.9. Triangulación entre hallazgos empíricos y referentes teóricos

Hallazgo cuantitativo	Hallazgo cualitativo	Referente teórico
Serie con tendencia y estacionalidad semestral; tres regímenes diferenciables (estabilidad, COVID, expansión).	El informante experto reconoce los patrones semestrales repetitivos y los tres regímenes históricos.	Hyndman & Athanasopoulos (2018): las series con tendencia y estacionalidad estable son apropiadas para Holt-Winters y SARIMA.
Método empírico actual asociado a alta variabilidad operativa.	Categoría C1: estimación por experiencia individual, sin procedimiento documentado.	Bird (2023): la planificación tradicional presenta limitaciones para capturar la complejidad de la matrícula.
Heterogeneidad alta entre carreras y necesidad de pronóstico desagregado por tiers.	Categoría C2: el informante atribuye los desajustes operativos a las carreras de crecimiento estructural.	Sghir et al. (2023): el pronóstico desagregado mejora la precisión institucional.
Descarte teórico y empírico de modelos no extrapolativos (Random Forest, XGBoost).	Categoría C3.1: el informante exige explicabilidad como condición de adopción.	Barredo Arrieta et al. (2020); James et al. (2021): la Explainable AI y la capacidad de extrapolación son condiciones de adopción.
Modelo final con MAPE 5,93 % (WF) y error 5,31 % (OOS 2026-P1) dentro del IC 95 %.	Categoría C3.2: expectativa de integración con tablero institucional para uso operativo.	Almalawi et al. (2024); Holmes & Tuomi (2022): la analítica predictiva integrada es condición de sostenibilidad institucional.

Fuente: elaboración propia.

La convergencia metodológica de los tres vértices refuerza la validez de los resultados. En particular, se corrobora que el problema de investigación planteado es real, vigente y reconocido por los actores institucionales. También se comprueba la pertinencia de las decisiones metodológicas adoptadas en el Capítulo II: el empleo de modelos paramétricos con tratamiento explícito de tendencia y estacionalidad, la desagregación por carrera, la clasificación por tiers de confiabilidad y la entrega del modelo a través de un tablero institucional.

Hay que destacar un punto de discusión metodológica de interés. La previa descartaría de algoritmos basados en árboles (Random Forest y XGBoost) por su falta de habilidad para extrapolar más allá del rango histórico se fundamenta teóricamente

en James et al. (2021) y Géron (2019). Los resultados cualitativos aportan un argumento institucional convergente: el informante experto prioriza la aplicabilidad frente a la complejidad algorítmica. La coincidencia entre el hallazgo técnico y la expectativa organizacional refuerza la decisión final y la legitima ante el tribunal evaluador.

La congruencia entre ambas validaciones independientes —MAPE walk-forward de 5,93 % y error fuera de muestra de 5,31 % frente al 2026-P1, período nunca observado durante el entrenamiento— es la evidencia técnica central del trabajo y demuestra que el modelo es estadísticamente robusto y operativamente útil para la PUCESA.

CONCLUSIONES

- Concluye, respecto del primer objetivo específico —teorizar la gestión de la matrícula estudiantil y el uso de modelos predictivos en la educación superior—, que la matrícula es un indicador estratégico que articula la asignación de carga docente, la apertura de paralelos, el uso de infraestructura y la planificación presupuestaria, y que su estimación mediante métodos empíricos es insuficiente frente a la complejidad del fenómeno. La revisión sistemática de fuentes arbitradas permitió establecer que el problema corresponde a una tarea de pronóstico de series temporales y delimitar seis algoritmos candidatos de tradiciones metodológicas diferenciadas.

- En cuanto al segundo objetivo específico —diagnosticar la situación actual de la gestión de la matrícula— se llega a la conclusión de que la institución no dispone de un procedimiento formal y documentado de proyección y que éste se basa en el criterio individual de los encargados. El diagnóstico cuantitativo, ejecutado sobre 31.875 registros estudiante–periodo depurados a 30.367 registros válidos (95,3 % de conservación) correspondientes a 22 periodos académicos (2015-P1 a 2025-P2), reveló una serie con tres regímenes claramente diferenciados: estabilidad entre 2015 y 2019 (1.175 a 1.315 estudiantes), caída por COVID-19 hasta el mínimo histórico de 806 estudiantes en 2021-P1, y expansión estructural posterior con crecimiento del 171,7 % hasta el máximo histórico de 3.142 estudiantes en 2025-P2, equivalente a un incremento acumulado del 138,9 % en el horizonte analizado.

- Respecto del tercer objetivo específico —implementar un modelo de predicción de matrícula mediante técnicas estadísticas y de aprendizaje automático— se concluye que el modelo desarrollado integra tres algoritmos seleccionados bajo un criterio doble objetivo definido a priori: desempeño empírico en validación walk-forward (MAPE inferior al 15 %) y capacidad de extrapolación. En este sentido, se seleccionaron los modelos SARIMA$(1,1,1)(1,0,1)_2$ (MAPE 11,61 %), Holt-Winters aditivo (11,83 %) y la Regresión de Huber (12,40 %); mientras que se descartaron XGBoost (14,54 %) y Random Forest (16,16 %) por su falta de capacidad de extrapolación, confirmando de manera empírica la hipótesis teórica del primer objetivo. La integración por mediana de los tres modelos elimina automáticamente las proyecciones extremas sin intervención manual, y el sistema entrega resultados desagregados por carrera con una clasificación por niveles de confia-

bilidad (tiers) que comunica de forma transparente la certeza diferenciada de cada predicción.

- Con relación al cuarto objetivo específico, validar el desempeño del modelo mediante pruebas de caja negra sobre datos históricos, se concluye que el modelo alcanza un desempeño de categoría excelente según la escala de Lewis (1982), corroborado por dos validaciones independientes y convergentes. La validación walk-forward en los últimos cinco periodos reales dio un MAPE de 5,93 %, mientras que la validación fuera de muestra en el periodo 2026-P1 —con 3.169 estudiantes reales que no participaron en ningún entrenamiento— tuvo un error de +5,31 % (predicción de 3.337 estudiantes), con el valor observado dentro del intervalo de confianza al 95 % construido por SARIMA.

- Con respecto al aporte tecnológico del prototipo se concluye que la solución es viable, escalable y usable en el contexto real de la institución. Su viabilidad se basa en una implementación íntegramente reproducible sobre software libre (Python, statsmodels y scikit-learn), que permite su replicación y mantenimiento sin dependencias propietarias; su escalabilidad se demuestra en el modelado desagregado de las 6 carreras con cobertura completa y la cobertura agregada de las 14 carreras detectadas, estructura que admite la incorporación progresiva de nuevos programas conforme vayan acumulando suficiente historia.

- Diseñar un modelo de predicción de matrícula para una institución de educación superior fue el objetivo general de la investigación, la cual fue satisfecha a través de un modelo de ensamble por mediana de SARIMA, Holt-Winters y Regresión de Huber, validado con 22 periodos de datos reales del sistema BANNER y con un error menor al 6 % en dos validaciones independientes (5,93 % en walk-forward y 5,31 % fuera de muestra). El trabajo muestra que es factible sustituir la estimación empírica de matrícula por una herramienta cuantitativa, interpretable y reproducible que proyecta el comportamiento institucional con precisión de categoría excelente —con una proyección de 3.450 estudiantes para el 2026-P2 y su intervalo de confianza respectivo— y aporta a la PUCESA un instrumento de planificación con base en evidencia que fortalece la asignación de recursos académicos, físicos y financieros y sienta las bases para su integración en un tablero analítico institucional.

RECOMENDACIONES

- A partir de los resultados obtenidos, se plantean las siguientes recomendaciones, las cuales se agrupan en tres ejes: operación y soporte del prototipo, infraestructura tecnológica y líneas de investigación futuras.

- • Se recomienda institucionalizar un procedimiento documentado de reentrenamiento del modelo al cierre de cada periodo académico, incorporando la matrícula real recién consolidada a la serie histórica, la validación walk-forward demostró que la precisión del modelo mejora conforme se acumulan observaciones del régimen vigente (el error descendió de 10,78 % a 3,32 % en los cinco periodos evaluados).

- • Se recomienda una política de respaldos automáticos de la base depurada y del cuaderno técnico reproducible con copias versionadas posteriores a cada extracción de BANNER para preservar la trazabilidad de las transformaciones del proceso ETL y garantizar la recuperación ante fallos.

- • Se recomienda mantener actualizadas las librerías de código abierto empleadas (pandas, statsmodels y scikit-learn) y aplicar los parches de seguridad correspondientes de forma periódica para conservar la compatibilidad del entorno reproducible.

- • Se recomienda proyectar un crecimiento del volumen de datos a mediano plazo, derivado de la apertura sostenida de nuevas carreras observada en la fase expansiva (+171,7 %), con el dimensionamiento del almacenamiento y del ancho de banda de extracción de BANNER para soportar la incorporación progresiva de programas y periodos sin degradar los tiempos de procesamiento.

- • Se sugiere la integración del modelo con un tablero analítico institucional, conectado de forma controlada a la fuente BANNER, conforme a la preferencia expresada en el diagnóstico, para que las proyecciones se actualicen automáticamente al cierre de cada periodo, y estén disponibles para la toma de decisiones.

BIBLIOGRAFÍA

Almalawi, A., Soh, B., Li, A., & Samra, H. E. (2024). Predictive models for educational purposes: A systematic review. *Big Data and Cognitive Computing, 8(12)*, 187. https://doi.org/10.3390/bdcc8120187

Arias Ortiz, E., Giambruno, C., Morduchowicz, A., & Pineda, B. (2024). *El estado de la educación en América Latina y el Caribe 2023*. Banco Interamericano de Desarrollo. https://doi.org/10.18235/0005515

Asamblea Nacional del Ecuador. (2021). *Ley Orgánica de Protección de Datos Personales*. Registro Oficial Suplemento 459 de 26 de mayo de 2021. https://www.telecomunicaciones.gob.ec/wp-content/uploads/2021/06/Ley-Organica-de-Datos-Personales.pdf

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bird, K. (2023). Predictive analytics in higher education: The promises and challenges of using machine learning to improve student success. *AIR Professional File*, (Fall 2023). https://doi.org/10.34315/apf1612023

Briggs, S. (2006). An exploratory study of the factors influencing undergraduate student choice: The case of higher education in Scotland. *Studies in Higher Education, 31(6)*, 705–722. https://doi.org/10.1080/03075070601004333

Brusnahan, A., Carrasco-Tenezaca, M., Bates, B. R., Roche, R., & Grijalva, M. J. (2022). Identifying health care access barriers in southern rural Ecuador. *International Journal for Equity in Health, 21(1)*, 55. https://doi.org/10.1186/s12939-022-01660-1

Campbell, J. P., & Oblinger, D. G. (2007). Academic analytics. *EDUCAUSE Quarterly, 30(4)*, 40–57.

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

Daniel, B. K. (Ed.). (2017). *Big data and learning analytics in higher education: Current theory and practice*. Springer. https://doi.org/10.1007/978-3-319-06520-5

Dietterich, T. G. (2000). Ensemble methods in machine learning. En J. Kittler & F. Roli (Eds.), *Multiple Classifier Systems* (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Fernández, M. (2021). *Gestión estratégica de la matrícula universitaria*. Editorial Universitaria.

Gartner. (s. f.). *Gartner's Analytics Ascendancy Model*. Gartner Research. Recuperado el 15 de octubre de 2025, de https://www.gartner.com/en/information-technology/glossary/analytics-ascendancy-model

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2ª ed.). O'Reilly Media.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2ª ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hernández-Sampieri, R., & Mendoza Torres, C. P. (2022). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta* (2ª ed.). McGraw-Hill Interamericana.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education, 57(4)*, 542–570. https://doi.org/10.1111/ejed.12533

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2ª ed.). OTexts. https://otexts.com/fpp2/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2ª ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Lewis, C. D. (1982). *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann.

Ministerio de Educación Nacional de Colombia. (2022, julio 1). El Ministerio de Educación Nacional pone a disposición la información estadística de educación superior a 2021. https://www.mineducacion.gov.co/portal/Noticias/Comunicados/411246:El-Ministerio-de-Educacion-Nacional-pone-a-disposicion-la-informacion-estadistica-de-educacion-superior-a-2021

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Perfumo, M. S., & Ares, M. V. (2023). Indicadores para la evaluación de la evolución de la matrícula en instituciones de educación superior. *593 Digital Publisher CEIT, 8(1)*, 24–38. https://doi.org/10.33386/593dp.2023.1.1052

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks, 16(3)*, 9–20. https://doi.org/10.24059/olj.v16i3.267

SENESCYT. (2023). *Informe sobre gestión y planificación de la educación superior en Ecuador 2023*. Secretaría de Educación Superior, Ciencia, Tecnología e Innovación. https://www.educacionsuperior.gob.ec

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and Information Technologies, 28*, 8299–8333. https://doi.org/10.1007/s10639-022-11536-0

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46(5)*, 30–32.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist, 57(10)*, 1510–1529. https://doi.org/10.1177/0002764213479366

UNESCO. (2021). *Reimagining our futures together: A new social contract for education*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000379707

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4ª ed.). Morgan Kaufmann. https://doi.org/10.1016/C2015-0-02071-8

ANEXOS

En esta sección se presentan los instrumentos de recolección de datos utilizados en la investigación y la referencia al cuaderno técnico reproducible que soporta el modelado. Los instrumentos se incluyen en su formato en blanco, conforme fueron validados por el director del Trabajo de Integración Curricular.

Anexo A. Ficha de Análisis Documental (FAD)

Instrumento estructurado empleado para homogeneizar la extracción y el registro de los datos cuantitativos provenientes de los reportes del sistema BANNER. Garantiza la trazabilidad y comparabilidad entre los veintidós periodos académicos analizados. Se diligencia una ficha por archivo-fuente procesado.

Tabla A. Ficha de análisis documental (FAD).

Campo	Registro
Sección 1. Identificación de la fuente	
Nombre del archivo	
Formato (Excel multipestaña / CSV / Asignaturas)	
Periodo BANNER (61 / 51 / 66 / 56)	
Año	
Semestre (P1 / P2)	
Fecha de extracción	
Sección 2. Variables cuantitativas extraídas	
Programa académico	
N.º de estudiantes inscritos	
N.º de estudiantes matriculados	
Modalidad de estudio	
Código BANNER de la carrera	
Sección 3. Variables derivadas	
Identificador secuencial de periodo (periodo_id)	
Familia normalizada de carrera (familia_carrera())	
Bandera: registro completo	
Bandera: periodo COVID (2020-P1 a 2021-P1)	
Bandera: programa nuevo	
Sección 4. Observaciones sobre la calidad del dato	
Registros faltantes	
Inconsistencias detectadas	
Valores atípicos	
Codificaciones especiales (Medicina 51/56; sufijos _R / _ Repotenciada)	
Validación cruzada con totales institucionales	
Sección 5. Validación	
Responsable de la extracción	
Fecha de validación	
Observaciones del director	

Fuente: elaboración propia, a partir de la matriz de operacionalización de variables (Capítulo II).

Anexo B. Guía de Entrevista Semiestructurada (GES)

Instrumento aplicado al Director Académico de PUCESA (Mg. Santiago Acurio) en su calidad de informante experto institucional, con el propósito de contextualizar los hallazgos cuantitativos, validar el problema de investigación y recoger expectativas sobre la implementación del modelo. Combina preguntas abiertas con ítems de percepción en escala de Likert de cinco puntos, organizados en cinco bloques temáticos alineados con los objetivos del estudio.

Datos del informante: cargo o función, escuela o dependencia, años de experiencia institucional, fecha de la entrevista y duración. Se aplica bajo consentimiento informado por escrito; la información se trata de forma confidencial, anonimizada y con fines exclusivamente académicos.

Bloque I. Procesos institucionales de matrícula

1. ¿Cómo describiría el proceso actual mediante el cual la institución estima la matrícula de cada nuevo periodo académico?

2. ¿Quiénes participan en esa estimación y con qué información se elabora?

Bloque II. Gestión y calidad de la información

3. ¿Qué sistemas o registros se utilizan como fuente de datos de matrícula?

4. ¿Qué dificultades de calidad o disponibilidad de información ha identificado?

Bloque III. Prácticas actuales de planificación y proyección

5. ¿Existe un procedimiento documentado para proyectar la matrícula? ¿Cuál es?

6. ¿Con qué anticipación se realiza la proyección frente al inicio del periodo?

7. ¿Qué ocurre cuando la matrícula real difiere de la proyectada?

Bloque IV. Factores institucionales que afectan la matrícula

8. ¿Qué factores internos o externos considera que más influyen en la variación de la matrícula?

9. ¿Cómo afectó la apertura de nuevas carreras (Medicina, Enfermería, Negocios Internacionales) a la planificación?

Fuente: elaboración propia, a partir de los bloques temáticos definidos en el Capítulo II.

Anexo B. Guía de Entrevista Semiestructurada (GES) (continuación)

Bloque V. Expectativas frente a un modelo predictivo

10. ¿Qué características debería tener un modelo de predicción de matrícula para que usted lo utilice?

11. ¿Cómo preferiría recibir los resultados del modelo (informe, tablero, archivo)?

Fuente: elaboración propia, a partir de los bloques temáticos definidos en el Capítulo II.

Ítems de percepción (escala de Likert de cinco puntos)

Marque el grado de acuerdo con cada afirmación: 1 = Totalmente en desacuerdo, 2 = En desacuerdo, 3 = Ni de acuerdo ni en desacuerdo, 4 = De acuerdo, 5 = Totalmente de acuerdo.

Afirmación	1	2	3	4	5
El método actual de estimación de matrícula es suficientemente preciso para la planificación.					
Existe un procedimiento institucional documentado para proyectar la matrícula.					
La información de matrícula disponible es confiable y oportuna.					
Los desajustes entre matrícula real y proyectada generan costos operativos relevantes.					
La apertura de nuevas carreras dificultó la planificación con el método actual.					
Un modelo predictivo basado en datos mejoraría la planificación académica.					
La explicabilidad del modelo es una condición indispensable para adoptarlo.					
Preferiría recibir las proyecciones integradas en un tablero institucional.					
Considero viable la adopción institucional de un modelo de pronóstico de matrícula.					

Fuente: elaboración propia, a partir de los bloques temáticos definidos en el Capítulo II.

Anexo C. Cuaderno técnico reproducible del modelo

La totalidad del procesamiento de datos, el modelado y las validaciones se implementaron en un cuaderno Jupyter reproducible (Defensa_Tesis_PUCESA_FINAL.ipynb), desarrollado en lenguaje Python sobre el entorno Google Colab. El cuaderno ejecuta de principio a fin sobre la serie real y reproduce íntegramente los resultados reportados en el Capítulo III. La Tabla D.1 sintetiza su estructura.

Tabla C.1. Estructura del cuaderno técnico reproducible

Sección	Contenido	Salida principal
ETL	Integración Excel + 6 CSV + 2 archivos 2026; deduplicación PIDM+periodo; familia_carrera()	30.367 registros válidos; 14 carreras
Exploración	Serie institucional, cobertura por carrera y descomposición orgánica/estructural	Figuras 3.1 – 3.3
Selección	Comparativa walk-forward y criterio doble (MAPE < 15 % y extrapolación)	Figura 3.5; Tabla 3.5
Modelado	Ajuste de SARIMA$(1,1,1)(1,0,1)_2$, Holt-Winters y Huber; ensemble por mediana	Figura 3.6; Tabla 3.6
Validación	Walk-forward de 5 periodos (MAPE 5,93 %) y out-of-sample 2026-P1 (5,31 %)	Figuras 3.7 – 3.9

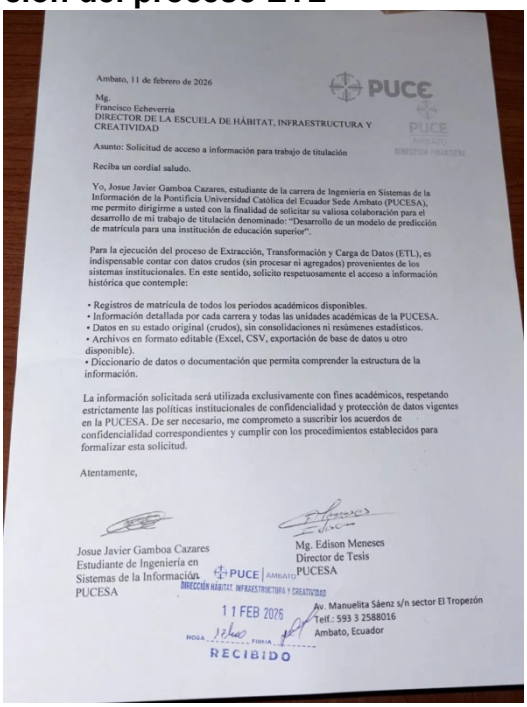
Fuente: elaboración propia. Librerías empleadas: pandas, numpy y openpyxl (procesamiento); statsmodels (SARIMA y Holt-Winters); scikit-learn (Regresión de Huber); matplotlib (visualización). El entorno, basado enteramente en software libre, garantiza la replicabilidad de los resultados por parte de terceros.

Repositorio del proyecto:

https://github.com/josueg1023434/DESARROLLO-DE-UN-MODELO-DE-PREDICCI-N-DE-MATR-CULA-PARA-UNA-INSTITUCI-N-DE-EDUCACI-N-SUPERIOR

El repositorio contiene el cuaderno ejecutable (Defensa_Tesis_PUCESA_FINAL.ipynb), los nueve archivos-fuente empleados (Excel histórico 2015–2020, seis CSV de inscritos 2020–2025 y dos archivos de asignaturas correspondientes al periodo 2026-P1) y el archivo README.md con la descripción del proyecto. El acceso es público y no requiere credenciales.

Anexo D. Solicitud formal de acceso a información institucional para la ejecución del proceso ETL



Ambato, 11 de febrero de 2026

Mg.
Francisco Echeverría
DIRECTOR DE LA ESCUELA DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

Asunto: Solicitud de acceso a información para trabajo de titulación

Reciba un cordial saludo.

Yo, Josue Javier Gamboa Cazares, estudiante de la carrera de Ingeniería en Sistemas de la Información de la Pontificia Universidad Católica del Ecuador Sede Ambato (PUCESA), me permito dirigirme a usted con la finalidad de solicitar su valiosa colaboración para el desarrollo de mi trabajo de titulación denominado: "Desarrollo de un modelo de predicción de matrícula para una institución de educación superior".

Para la ejecución del proceso de Extracción, Transformación y Carga de Datos (ETL), es indispensable contar con datos crudos (sin procesar ni agregados) provenientes de los sistemas institucionales. En este sentido, solicito respetuosamente el acceso a información histórica que contemple:

- Registros de matrícula de todos los periodos académicos disponibles.
- Información detallada por cada carrera y todas las unidades académicas de la PUCESA.
- Datos en su estado original (crudos), sin consolidaciones ni resúmenes estadísticos.
- Archivos en formato editable (Excel, CSV, exportación de base de datos u otro disponible).
- Diccionario de datos o documentación que permita comprender la estructura de la información.

La información solicitada será utilizada exclusivamente con fines académicos, respetando estrictamente las políticas institucionales de confidencialidad y protección de datos vigentes en la PUCESA. De ser necesario, me comprometo a suscribir los acuerdos de confidencialidad correspondientes y cumplir con los procedimientos establecidos para formalizar esta solicitud.

Atentamente,

Josue Javier Gamboa Cazares
Estudiante de Ingeniería en Sistemas de la Información
PUCESA

Mg. Edison Meneses
Director de Tesis
PUCESA

PUCE | AMBATO
DIRECCIÓN HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

Av. Manuelita Sáenz s/n sector El Tropezón
Telf.: 593 3 2588016
Ambato, Ecuador

11 FEB 2026
HORA
FIRMA
RECIBIDO

Figura D.1. Oficio dirigido al Mg. Francisco Echeverría, director de la Escuela de Hábitat, Infraestructura y Creatividad de la PUCESA, mediante el cual se formaliza el requerimiento de acceso a los registros históricos de matrícula del sistema BANNER.
Fuente: elaboración propia