ANÁLISIS PREDICTIVO DEL CRECIMIENTO POBLACIONAL DE LA UNIVERSIDAD NACIONAL DE COLOMBIA, SEDE BOGOTÁ.

JUAN FELIPE FORERO BOCANEGRA
DANIEL ANDRÉS PEDRAZA ROMERO

UNIVERSIDAD EL BOSQUE PROGRAMA DE INGENIERÍA DE SISTEMAS
FACULTAD DE INGENIERÍA
Bogotá, 2022

ANÁLISIS PREDICTIVO DEL CRECIMIENTO POBLACIONAL DE LA UNIVERSIDAD NACIONAL DE COLOMBIA, SEDE BOGOTÁ.

JUAN FELIPE FORERO BOCANEGRA
DANIEL ANDRÉS PEDRAZA ROMERO

Desarrollo Tecnológico presentado como requisito para optar al título de

INGENIERO DE SISTEMAS

Director
CARLOS DELGADO ROMÁN
Ingeniero de Sistemas

UNIVERSIDAD EL BOSQUE
PROGRAMA DE INGENIERÍA DE SISTEMAS
FACULTAD DE INGENIERÍA
Bogotá, 2022

Contenido

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

La Oficina de Planeación y Estadística (OPE) de la Universidad Nacional de Colombia es la encargada de implementar las políticas de planeación y desarrollo institucional de la Sede Bogotá. Actualmente, los integrantes de la OPE han llevado a cabo un trabajo de recopilación, estructuración y descripción de la información que se consolida en tableros interactivos de Tableau y la Plataforma de Registro de Informes de Gestión (PRIG). Sin embargo, no se ha realizado ningún trabajo de analítica predictiva con técnicas propias de sistemas inteligentes. Este trabajo se propone así dar el primer paso en este sentido al realizar modelos predictivos de las tendencias de comportamiento del crecimiento poblacional de matriculados, docentes, graduados y administrativos a partir de las bases de datos de la OPE mediante algoritmos de machine learning con el fin de apoyar la planeación estratégica institucional. El proyecto se desarrolló con la metodología especializada en proyectos de análisis de datos ASUM-DM y la investigación cualitativa, la cual permitió medir la percepción de utilidad de los modelos. A partir del Ciclo de Transferencia Tecnológica (CTT) y el Modelo Biopsicosocial y Cultural de la Universidad El Bosque, se logró transformar los hábitos de los integrantes de la OPE en el uso de herramientas de analítica predictiva. Los modelos se elaboraron con técnicas de *auto machine learning* y el algoritmo utilizado para todos los conjuntos de datos fue el *Random Forest Regressor* con un 90 % de R2. La importancia de los modelos radica en la optimización de recursos físicos, humanos y financieros y poder prever la demanda de los servicios ofrecidos. Esto afecta las dimensiones económicas, sociales y culturales del entorno universitario ya que facilita la planeación estratégica de la universidad.

INTRODUCCIÓN

La toma de decisiones basada en datos se ha convertido en un factor fundamental para cualquier tipo de organización ya que permite realizar acciones de mejora de sus procesos. Mediante la aplicación de algoritmos de machine learning, se logran identificar patrones ocultos en los datos que facilitan decisiones de negocio mucho más acertadas de cara a una mayor proyección empresarial y ventaja competitiva. De acuerdo con esto, las empresas e instituciones deben empezar a enfocar sus estrategias comerciales y organizacionales hacia un entorno basado en datos.

Toda organización en el ejercicio de sus labores diarias genera información, la cual puede ser recolectada desde múltiples fuentes asociadas a un proceso de negocio específico. Sin embargo, se debe tener en cuenta que no todos los datos recolectados son relevantes para los objetivos de una organización, es por esta razón que resulta necesario someter a un proceso de limpieza y transformación el conjunto de datos con el fin de sacarles el máximo provecho. Esto permitirá iniciar el análisis que busca extraer información significativa ya sea mediante métodos estadísticos o técnicas de aprendizaje máquina y cuyos resultados pueden generar valor agregado al negocio. Lo anterior configura lo que se denomina el ciclo de vida de los datos y es el punto de partida en todo proyecto de análisis.

En tal sentido, la Universidad Nacional de Colombia, a través de la Oficina de Planeación y Estadística (OPE) viene trabajando activamente en el análisis de un histórico de más de 10 años de datos sobre los cuales buscan determinar el comportamiento del crecimiento de la población universitaria de cara a mejorar la planeación institucional, sin embargo, actualmente el análisis se realiza de manera manual y puramente descriptiva, lo cual dificulta contemplar todo el conjunto de datos para determinar y predecir el crecimiento de la población de una manera más acertada. Este documento aborda esta problemática y propone una solución de ingeniería basada en la aplicación de algoritmos de *machine learning,* detallando el proceso necesario de acuerdo con el marco metodológico ASUM-DM. Como resultado, se obtuvieron modelos con una precisión de más del 90% con el algoritmo *Random Forest Regressor* al emplear técnicas de *auto machine learning*. De acuerdo con el ciclo de transferencia tecnológica y modelo biopsicosocial y cultural de la Universidad del Bosque, se dió el primer paso en transformar los hábitos de los agentes de la OPE en el uso de la analítica predictiva en la planeación

estratégica institucional. Esto es importante porque estas técnicas permiten reducir la incertidumbre, evitar riesgos y aprovechar oportunidades.

La estructura del documento sigue los procesos del ciclo de transferencia tecnológica en la medida en que, en el primer capítulo, se identifica la problemática desde el modelo biopsicosocial y cultural, en el segundo, se plantea una solución, en el tercero, se establece el marco referencial con sus correspondientes antecedentes, en el cuarto, se evidencia todo el desarrollo metodológico a partir de ASUM-DM, en el quinto, se analizan los resultados, en el sexto, se concluye y, en el séptimo, se exponen las lecciones aprendidas.

1. PROBLEMA

Según los lineamientos del programa de Ingeniería de Sistemas de la Universidad El Bosque, el ciclo de transferencia de tecnología permite estructurar el proyecto de grado de forma que cumpla con los estándares tanto académicos como empresariales del artefacto propuesto. Así, el primer paso es identificar el problema, describiendo el contexto a través de entrevistas u observaciones de campo a partir del modelo biopsicosocial y cultural de la universidad. En el presente capítulo se identificará la problemática a resolver a partir de estas herramientas.

1.1 Descripción del contexto a intervenir

La Oficina de Planeación y Estadística (OPE) de la Universidad Nacional de Colombia, Sede Bogotá, tiene como objetivo implementar las políticas y reglamentaciones de los procesos de análisis del desarrollo institucional. Es la encargada de realizar estudios que apoyen la toma de decisiones en la gestión de la sede, la divulgación de indicadores y la rendición pública de cuentas. Se cuenta con un histórico de más de diez años de información de diferentes bases de datos institucionales con las cuales se han realizado análisis descriptivos que se materializan en reportes interactivos de acceso abierto disponibles en la página oficial. Los principales temas abordados son: 1) calidad académica, que cuenta con tableros sobre los programas académicos, la caracterización de los docentes y los administrativos, la infraestructura y las pruebas Saber Pro; 2) caracterización estudiantil, que reúne los datos sobre los matriculados y los graduados; 3) indicadores académicos, que sintetizan las registros de los aspirantes admitidos a pregrados y posgrado, las convocatorias estudiantiles auxiliares y la deserción; 4) investigación y extensión, que integra los registros sobre los grupos, las modalidades y los eventos de investigación, los proyectos de extensión, la productividad investigativa y las revistas indexadas; y, por último, 5) la movilidad estudiantil y docente, que reúne la información sobre el mismo tema. El objetivo de la OPE, a largo plazo, es crear un sistema de información integrado para la sede de Bogotá que apoye la planeación institucional.

Según el asesor de la OPE, para la elaboración de los reportes interactivos se analiza la información mediante archivos Excel, con los cuales se hace un análisis parcial dada la cantidad de datos que se tienen, por lo tanto, en la medida que son generados los datos del semestre inmediatamente anterior, estos datos son analizados de manera segregada y pasan a alimentar los tableros interactivos. Sobre el resultado de ese análisis, se toman decisiones para la ejecución de plan institucional.

1.2 Análisis del contexto desde el modelo biopsicosocial y cultural

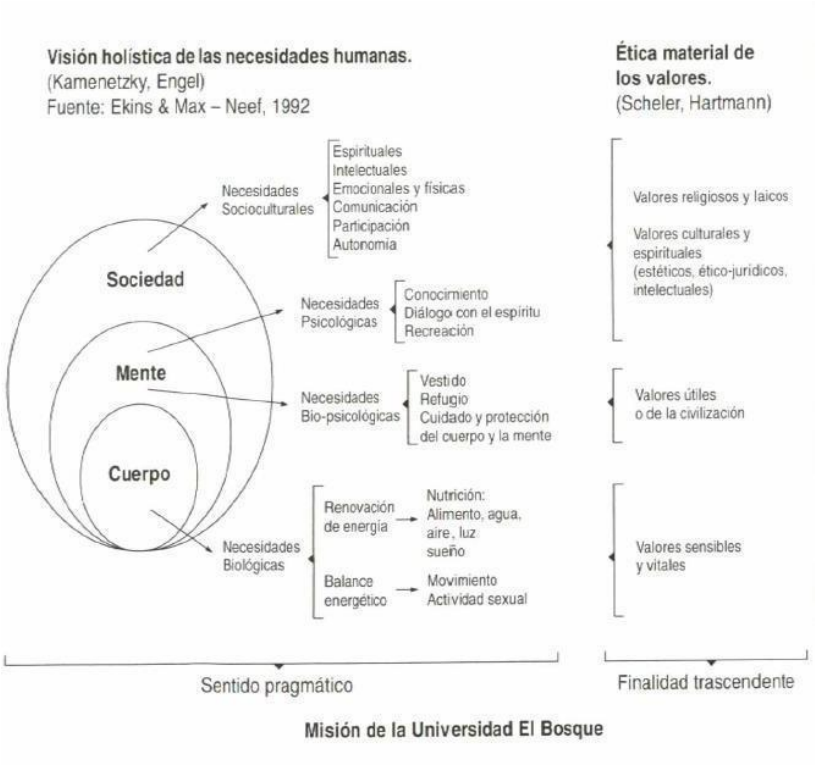
La Escuela Colombiana de Medicina surge como una fundación sin ánimo de lucro, que intenta dar respuesta a la crisis de la educación médica nacional mediante un enfoque alternativo. Este parte de un concepto antropológico del ser humano "en un proceso dinámico de salud-enfermedad, que considera no solo sus aspectos biológicos sino también sus aspectos psicológicos y sociales". [1, p. 204] La Universidad El Bosque asume su compromiso de educación con el imperativo supremo de la promoción de la dignidad del ser humano en su integridad. Su esfuerzo se concentra en ofrecer las condiciones propias para facilitar el desarrollo de valores éticos, morales, estéticos y científicos en una cultura de la vida, su calidad y su sentido. [1, p. 206]

El modelo biopsicosocial y cultural se centra entonces en una visión integral de las necesidades humanas, según se exponen en la figura 1. Por su parte, el modelo para ingeniería se construye a partir de cuatro elementos principales: creencias o cultura, hábitos, medio y artefactos. Estos están articulados desde una perspectiva sistémica en la que el cambio de alguno significa un cambio en el resto. [2, p. 2] Las creencias o la cultura se enmarcan como un tejido de normas a través de las cuales un grupo o una comunidad regula la relación entre sus miembros y con el medio. Una norma es una guía de acción considerada adecuada o justa. El medio es el lugar en el cual se ubica el grupo y del cual obtiene lo necesario para su subsistencia. El hábito es un conjunto de acciones repetitivas que realiza el conjunto de las personas. El artefacto, por último, es un objeto diseñado para cumplir una función específica. [2, p. 4]

La Ingeniería de Sistemas desarrolla entonces *artefactos* de software que permiten la interpretación y el análisis de su influencia en los demás aspectos del modelo, es decir, en este artefacto confluyen las demás instancias ya que transforma e influye en la cultura, el medio y los hábitos. No obstante, este está delimitado también por las demás del mismo modo por el carácter sistémico del modelo. [2, p. 4]

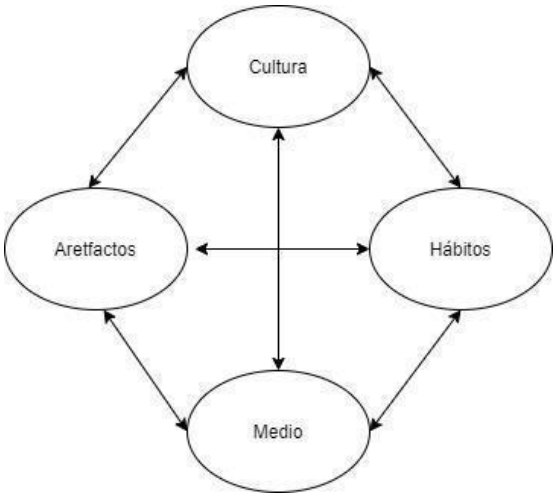
En el siguiente apartado, se identifica la problemática en el contexto de la Oficina de Planeación y Estadística (OPE) de la Universidad Nacional de Colombia a partir del modelo biopsicosocial y cultural.

*Figura SEQ Figura * ARABIC 1. Modelo Biopsicosocial y cultural*



Fuente: Tomado de [1, p. 2016].

*Figura SEQ Figura * ARABIC 2. Modelo BPSC para ingeniería*



Fuente: Tomado de [2, p. 4]

Análisis del contexto del artefacto

La Oficina de Planeación y Estadística (OPE) de la Universidad Nacional de Colombia (UNAL) es la encargada de coordinar e integrar los procesos de planeación de la sede de Bogotá y de asesorar a las diferentes dependencias en la implementación del Plan Estratégico Institucional, el Plan Global de Desarrollo y el Plan de Acción Institucional. A grandes rasgos, sus principales funciones consisten en implementar las políticas y reglamentaciones en materia de planeación y estadística; participar en los procesos de planeación de la sede; producir, procesar y divulgar indicadores y estadísticas; realizar estudios para la toma de decisiones y participar en los procesos de rendición pública de cuentas y en la elaboración de planes de desarrollo institucional. [3]

De esta forma, se identifica el *medio* para este contexto como la Universidad Nacional de Colombia dado que es la que enmarca y da los lineamientos necesarios para el funcionamiento de la OPE dentro de una *cultura* institucional bien definida por el Plan Estratégico Institucional, el Plan Global de Desarrollo y el Plan de Acción Institucional. Los *hábitos*, por su parte, corresponden al proceso de análisis descriptivo sobre los datos y la toma de decisiones sobre los resultados para cumplir los objetivos organizacionales planteados. En cuanto a los *artefactos* de interés, encontramos los tableros interactivos que permiten a la comunidad universitaria y a la ciudadanía, en general, consultar información detallada e interrelacionada sobre temas y cifras de la sede de Bogotá. Actualmente, se tiene acceso a los siguientes tableros, que son de acceso abierto en la página oficial de la OPE:

- Descripción histórica y actual de los programas académicos.
- Caracterización docente del distrito y del Sistema de Universidades Estatales u Oficiales (SUE) entre 2007 y 2017.
- La caracterización del personal administrativo y su distribución por nivel, sexo y formación.
- Evolución de matriculados en Instituciones de Educación Superior (IES) de acuerdo con la información suministrada por el Sistema Nacional de Información de la Educación Superior (SNIES).
- La caracterización de los matriculados en la Sede de Bogotá a partir de 2009.
- Los graduados en el distrito, los graduados SUE distrital y de la Sede Bogotá entre 2001 y 2017.

- Caracterización de los docentes a partir de 2008.
- Caracterización de aspirantes y admitidos a pregrado de la Sede Bogotá, entre 2011 y 2021.
- Caracterización de aspirantes y admitidos a posgrado de la Sede Bogotá, entre 2011 y 2021.
- Convocatoria de estudiantes auxiliares.
- Deserción estudiantil en universidades del distrito y en la Sede de Bogotá entre 2001 y 2020 según el Sistema para la Prevención y Análisis de la Deserción en la Instituciones de Educación Superior.
- Grupos de Investigación con participación de la UNAL a partir de 2013.
- Productividad investigativa desde 2009.
- Movilidad estudiantil entrante y saliente en la sede entre 2018 y 2019.
- Movilidad docente entrante y saliente en la sede entre 2018 y 2019.

Estas herramientas facilitan, entonces, la rendición de cuentas públicas y la construcción de históricos de estadísticas para la planeación estratégica. Su impacto en el medio, las creencias y los hábitos, tanto de la OPE como de la UNAL, en general, es alto, debido a que el Plan Estratégico Institucional y el Plan de Acción Institucional para cada gobierno se realizan con base en los registros de los informes de gestión que se recopilan en la Plataforma de Registro de Informes de Gestión (PRIG) y los tableros interactivos. Por su parte, el PRIG tiene como propósito la sistematización del proceso de reporte anual de logros, dificultades, mejoras y acciones correctivas de las diferentes dependencias de la universidad. Sus objetivos son consolidar las cifras de las actividades de cada dependencia y área; definir los indicadores; generar boletines de divulgación que apoye la toma de decisiones y permitir la trazabilidad de la gestión institucional a lo largo del tiempo. [4] Así, los reportes interactivos y los informes de gestión son el insumo para una planeación basada en datos.

De acuerdo con el asesor de la OPE, existen dos necesidades fundamentales para el análisis de esta información: la visualización de los datos obtenidos en el PRIG y el análisis del crecimiento poblacional en matriculados, docentes, graduados y administrativos. El concepto de crecimiento poblacional se puede entender como el cambio de la población en un tiempo definido.

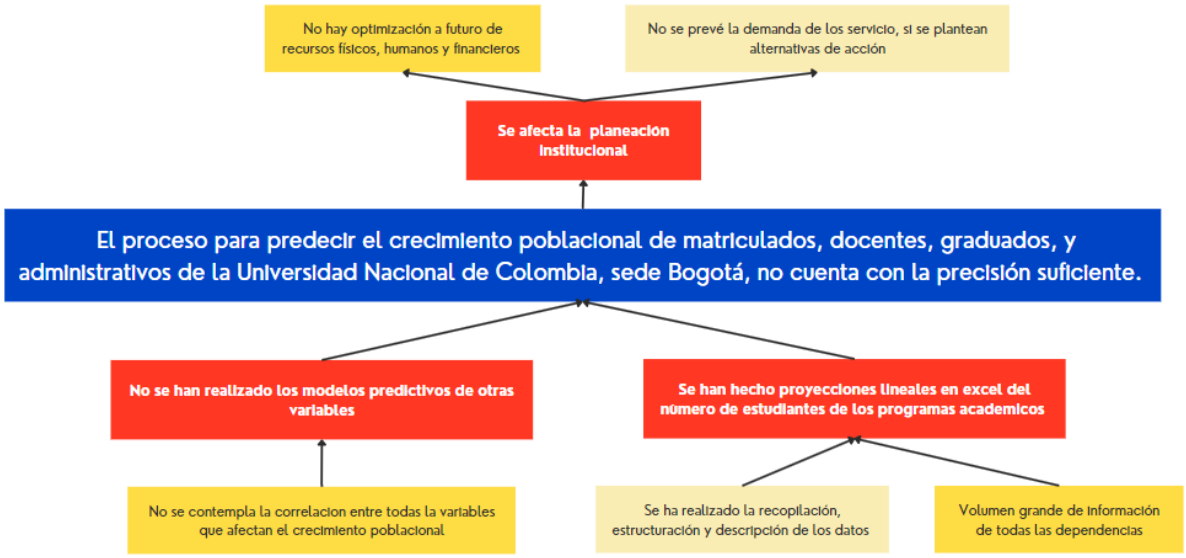
1.3 Identificación y descripción de la problemática

Hasta la fecha se han realizado análisis descriptivos de la información disponible de la sede Bogotá. Estos se materializan en los reportes interactivos descritos y en los informes de gestión recopilados en el PRIG. Se ha logrado recopilar, estructurar y describir la información, sin embargo, para la UNAL no es posible determinar cuál será el comportamiento de las cifras a futuro, debido a que los análisis realizados son de tipo descriptivo y no es posible hacer predicciones.

A partir del modelo biopsicosocial (ver anexo 1) ya se identificó el medio como la Universidad Nacional de Colombia, sede Bogotá. Los actores son los analistas de datos de la OPE cuyos hábitos consisten en recopilar y estructurar la información. Este trabajo se consolida en los reportes interactivos de Tableau y los informes de gestión, los cuales afectan las dimensiones económicas y sociales del entorno universitario ya que es con base en estos que se elaboran los planes de gestión. Las creencias de los funcionarios es que es indispensable crear un sistema de información unificado para todas las dependencias de modo que se pueda explotar al máximo las cifras disponibles. Esto afecta las dimensiones culturales, sociales y económicas dado que una mejor planeación mejora la calidad de vida de la comunidad universitaria.

No se conoce el crecimiento poblacional de matriculados, graduados, docentes y administrativos a futuro, lo que afecta la administración de espacios, el desarrollo de proyectos y la generación de programas. Como solución parcial a la fecha, la OPE ha hecho proyecciones lineales en Excel del número de estudiantes de los diferentes programas académicos. Una de las desventajas de este enfoque es que el índice de correlación no es muy preciso y no se está recogiendo las características poblacionales de los estudiantes.

Figura 3. Árbol de problema



En la figura 3 se puede apreciar el árbol de problema, en el cual se establecen las causas (o la raíz) y las consecuencias (o las hojas) de la problemática. Así, se cuentan con un volumen grande de información de todas las dependencias de la Universidad Nacional y se ha realizado un trabajo de recopilación, estructuración y descripción de los datos. Sin embargo, por un lado, no contemplar el histórico de datos que se tienen en el análisis, implica un alto riesgo en la planeación y, por el otro, no tener una buena precisión en la medición futura del crecimiento poblacional, puede desembocar una gestión institucional inadecuada con una asignación presupuestal desfasada, una gestión pobre de los espacios físicos y proyectos académicos que no se alinean a las demandas del mercado.

2. SOLUCIÓN DE INGENIERÍA

La segunda etapa del ciclo de transferencia tecnológica de ingeniería es la formulación del problema de acuerdo con la necesidad, la oportunidad o el reto identificado a partir del modelo biopsicosocial y cultural. En este apartado se definen los objetivos generales y específicos, se describe la solución planteada y se establece la metodología de desarrollo y los resultados esperados por el cliente.

2.1 Objetivos: general y específicos

Objetivo general

Desarrollar modelos predictivos de las tendencias de comportamiento del crecimiento poblacional de matriculados, docentes, graduados y administrativos de la Universidad Nacional de Colombia, sede Bogotá, a partir de las bases de datos de la Oficina de Planeación y Estadística (OPE) mediante algoritmos de *machine learning* para apoyar la planeación estratégica institucional.

Objetivos específicos

1. Determinar los requerimientos de información y las fuentes de datos necesarias que agreguen valor en el análisis sobre las tendencias de comportamiento.
2. Definir y aplicar los algoritmos que permitan predecir el crecimiento poblacional siguiendo el ciclo de vida de los datos: planificación, recolección, control de calidad, almacenamiento, integración y análisis.
3. Validar el impacto del modelo en la planeación estratégica de la OPE respecto a la percepción de utilidad de las predicciones.

2.2 Descripción de la solución y resultados esperados

En este proyecto se propone la elaboración de modelos predictivos de las tendencias de comportamiento poblacional de matriculados, docentes, graduados y administrativos de la sede Bogotá de la Universidad Nacional de Colombia. Los modelos se realizan con técnicas de *machine learning* o aprendizaje de máquina, cuya precisión debe ser mayor al 80 %. Los resultados deben ser visibles en tableros disponibles en la web que permitan discriminar las diferentes características poblacionales para facilitar así su socialización. Adicionalmente, los modelos deben ser susceptibles de actualización con el ingreso de nuevos datos de entrada.

2.3 Análisis de la solución desde el modelo biopsicosocial y cultural

El medio del modelo ya fue definido como la Universidad Nacional de Colombia, específicamente, la Oficina de Planeación y Estadística. Los actores de la solución serían los analistas de datos cuya cultura organizacional está enfocada en crear un sistema de información integrado de la sede Bogotá que apoye la planeación. Se propone crear el hábito de realizar proyectos de analítica predictiva con el objetivo de disminuir la incertidumbre en la toma de decisiones. Esto afecta las dimensiones económicas, sociales y culturales del entorno universitario ya que facilita el desarrollo de proyectos, la gestión de los programas académicos, la asignación presupuestal y la administración de espacios físicos (ver anexo 2).

2.4 Tabla de entregables

En la tabla 1 se muestran los entregables generados por cada fase de acuerdo con la metodología de analítica de datos ASUM-DM; adicionalmente, se muestran los criterios de aceptación para cada entregable que configuran el cumplimiento de cada una etapa tras la aprobación por parte del beneficiario.

Tabla 1. Entregables del proyecto

Artefactos		
Fase	**Entregable**	**Criterio de aceptación**
Análisis	Acta de constitución	• Plan de costos y riesgos aprobado por el beneficiario.
	Plan de costos y riesgos	• Detalle del análisis técnico aprobado por el beneficiario
Diseño	Reporte de preparación de datos	• Detalle del proceso de limpieza y transformación de datos. • Detalle del análisis descriptivo y exploratorio sobre el conjunto de datos.
Construcción	Reporte de construcción del modelo	• Detalle del proceso de elección del algoritmo y el modelo construido
Evaluación	Reporte de evaluación del modelo	• Detalle de los resultados obtenidos tras evaluación del modelo.
Despliegue	Reporte de transferencia de conocimiento	• Manual de usuario • Manual técnico
Cierre	Acta de cierre	• Acta de cierre del proyecto

2.5 Variable a medir

El concepto de crecimiento poblacional se entiende como el cambio o la variación de la población en un periodo de tiempo determinado. En este caso, se cuenta con los registros desde 2009 hasta la fecha de matriculados, docentes, graduados y administrativos. Hasta el momento solo se han realizado proyecciones lineales en Excel del número de estudiantes por programa académico con una precisión baja que no permite realizar un análisis multivariado del conjunto de características del grupo

estudiado. Se espera que, con las técnicas de aprendizaje automático, los modelos se ajusten al conjunto de los datos y generen predicciones más acertadas, teniendo en cuenta las variables pertinentes y las características intrínsecas del grupo estudiado.

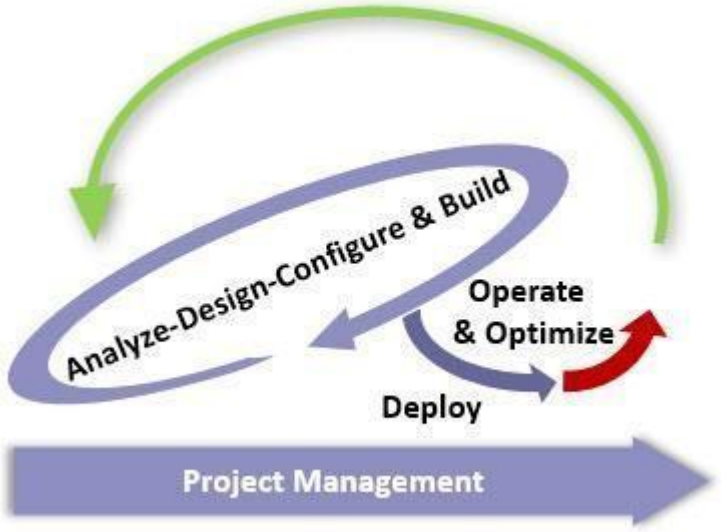
En el marco del ciclo de transferencia tecnológica, se realizará una validación académica en un ambiente controlado del artefacto para garantizar su calidad. Siguiendo la metodología de desarrollo de proyectos de analítica de datos ASUM-DM, se realizarán las etapas de un proyecto de analítica de datos. La solución de ingeniería se construirá de la mano con el beneficiario de acuerdo con el cronograma planteado con el fin de realizar la validación estática, que se concretará en cada uno de los entregables de las fases. En cuanto a la validación dinámica del artefacto en un contexto real, que tiene como fin medir el impacto de la solución a partir del objetivo general planteado, se empleará la investigación cualitativa y la metodología de encuesta para medir la percepción de utilidad de los modelos predictivos por parte de los agentes involucrados en la planeación institucional de la universidad.

2.6 Metodología

La metodología ASUM-DM (figura 4) propone un proceso iterativo orientado a proyectos de analítica de datos. Los componentes claves son:

- **Entendimiento del negocio**. En esta etapa se busca obtener la mayor información acerca del negocio con relación a la problemática específica y los beneficios que se esperan obtener para la organización.
- **Preparación de los datos**. Se describe todo el proceso de recolección, clasificación, exploración y limpieza, lo que permitirá conocer su volumen y su calidad.
- **Construcción del modelo**. En este punto se determina qué técnicas de aprendizaje, clasificación y predicción son las que mejor se ajustan a las necesidades de la organización.
- **Evaluación del modelo**. Se realiza una valoración de los resultados del modelo en relación con los criterios de aceptación establecidos.
- **Despliegue.** La solución pasa a un ambiente productivo en el que el usuario podrá usar el artefacto.

Figura 4. Metodología ASUM -DM



Fuente: Tomado de [5]

Esta metodología propone a su vez un componente de gestión de proyectos transversal a las fases mencionadas anteriormente, en el cual se establece la planificación, el monitoreo y la optimización del proyecto en sí y del producto a entregar. Teniendo en cuenta la problemática descrita anteriormente, la metodología ASUM- DM es la que mejor se adapta al proceso requerido para lograr alcanzar el objetivo de una planeación estratégica basada en análisis de datos.

ASUM-DM, nace como una extensión de la metodología CRISP-DM, es decir que adopta gran parte de los componentes que forman parte de esta. La elección de ASUM-DM, fundamentalmente, se basa en el detalle que brinda para la ejecución de cada una de las fases [20] y, además, incorpora, como parte de la ejecución del proyecto, el componente de implementación, que no fue considerado en la metodología CRISP-DM. Por otra parte, y en comparación con la metodología TDSP de Microsoft, ASUM-DM brinda un mayor detalle en relación con las responsabilidades de cada uno de los roles que intervienen.

De acuerdo con la ruta de implementación de ASUM-DM, la metodología se divide en entendimiento de negocio, entendimiento de los datos, preparación de estos y construcción y evaluación del modelo, lo cual representan elementos importantes para el logro de los objetivos uno y dos de este proyecto. Sin embargo, para el tercer objetivo, la metodología ASUM-DM no brinda componentes que permitan

validar el impacto del artefacto. Por tal motivo, se realizará una aproximación de investigación cualitativa en la cual se empleará la metodología de encuesta para medir la percepción de utilidad de los modelos en la planeación estratégica.

2.6.1 Estructura de desglose de trabajo

El plan de trabajo se organizó en nueve paquetes o bloques que se corresponden con el ciclo de implementación de proyectos de analítica de datos (ver anexo 3). En el primer bloque, se establecen las actividades de planificación y gestión en las cuales se lleva a cabo un análisis técnico que permite tener una visión global del negocio y plantear los requerimientos de alto nivel. Este análisis cuenta con un Plan de riesgos, un Plan de costos y un Plan de calidad. Este proceso culmina con un acta de constitución en el que se define el propósito y la justificación del proyecto, los entregables, los requerimientos del producto y del proyecto en sí, los objetivos, las restricciones, los riesgos, el cronograma, el presupuesto y los requisitos de aprobación. El acta de constitución debe ir firmada por todos los interesados, en este caso el patrocinador y asesor de la OPE y los autores del presente estudio.

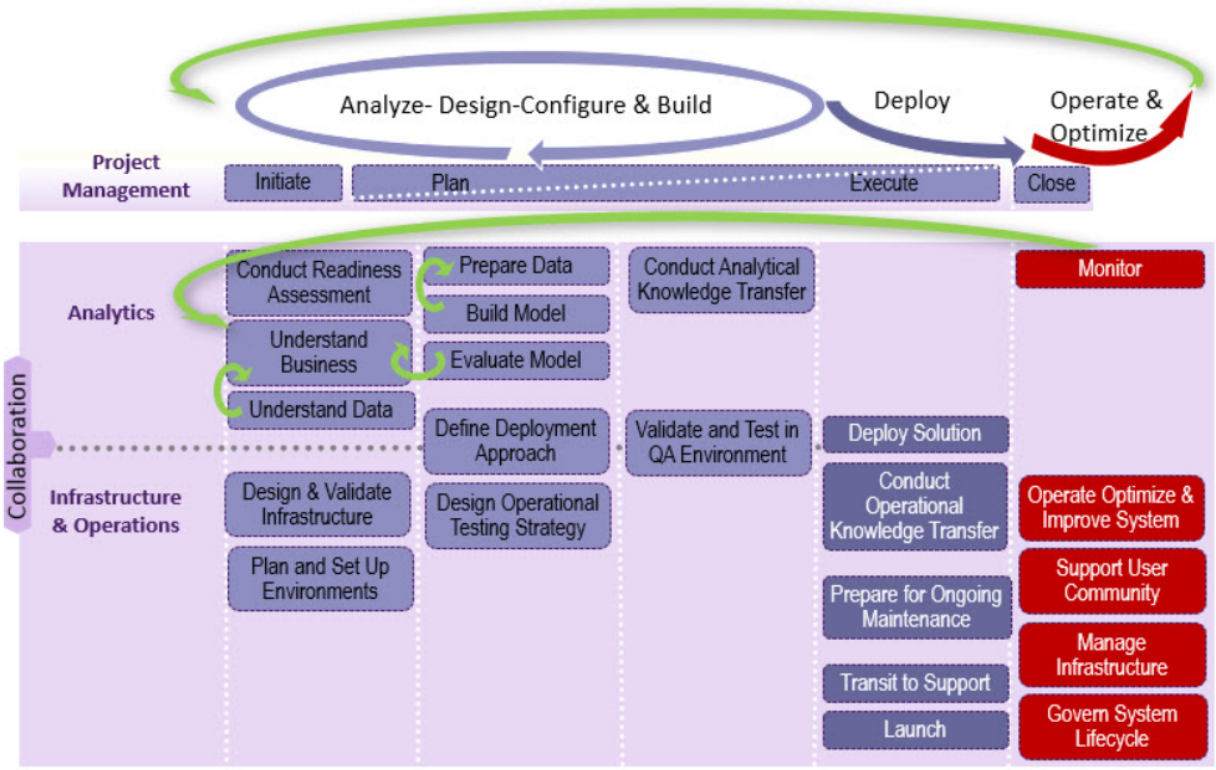
En el segundo bloque de trabajo se realiza un análisis del negocio que permite levantar los requerimientos funcionales y no funcionales y los límites y restricciones. En el tercer bloque se realiza el diseño de la solución que parte de la selección de la información a partir de la variable a medir para luego realizar la limpieza, la transformación, la integración y la normalización de los datos. Esta actividad quedará registrada en un reporte de preparación de los datos que se le entregará al cliente.

En el cuarto bloque se seleccionan las técnicas de aprendizaje automático que más se adecuen a los objetivos del proyecto. En esta fase se realiza también el plan de pruebas y la construcción del modelo. En el quinto bloque se evalúan los resultados hasta el momento y se establece el plan de transferencia de conocimiento y el plan de despliegue. En el sexto se ejecuta el plan de pruebas para el aseguramiento de la calidad y se corrigen los errores hasta que los modelos sean aprobados por el cliente. En el séptimo se ejecuta el plan de despliegue, en el cual se establece un ambiente de producción y se le indica al usuario cómo usar el artefacto. En el octavo se realiza la validación dinámica en la que se mide el impacto de la solución en la planeación estratégica institucional. Por último, se da el cierre del proyecto con el informe y el acta correspondientes.

2.6.2 Aplicación de la metodología

La estructura de desglose de trabajo sigue la ruta de implementación de ASUM-DM (ver figura 5) según la cual es indispensable realizar un proceso iterativo de mejoramiento continuo. Así, la gestión del proyecto es transversal a las actividades de análisis, diseño, construcción, validación y cierre, realizando siempre ciclos de monitoreo constantes que permitan optimizar los resultados, revaluar los objetivos y, en última instancia, asegurar el éxito de la solución.

Figura 5.Ruta de implementación de ASUM-DM



Fuente: Tomado de [5]

Como parte del componente de proyecto propuesto en la metodología se estimó el alcance y esfuerzo requerido para la ejecución del proyecto planteado, dando como resultado el plan de costos y riesgos (ver anexo 5), lo anterior es un insumo importante para pasar a la fase de entendimiento de negocio, donde el asesor de la OPE hace entrega de un conjunto de datos inicial. Sobre este conjunto de datos se procede a ejecutar la fase de preparación de datos.

2.6.3 Cronograma

En el anexo 4 se puede ver el cronograma inicial planteado, sin embargo, luego de varias conversaciones con el cliente, se decidió no hacer el proceso para los 4 data sets mencionados inicialmente, sino hacer el proceso para un solo dataset, concretamente el de matriculados, con el fin de tener prototipo funcional que dé claridad sobre todo el proceso y, a partir de esto, generar una plantilla o modelo de trabajo base que sea replicado durante el periodo intersemestral en los data sets restantes.

2.7 Acuerdo con el cliente

El acta de constitución del proyecto (ver anexo 6) establece el propósito y la justificación del mismo y define los entregables con sus respectivos requerimientos de producto. Así mismo, se plantean los objetivos generales y específicos, los supuestos y restricciones, los riesgos, el cronograma, el presupuesto, los encargados y los requisitos de aprobación para establecer el cierre. Esta acta constituye el resultado del análisis técnico.

Es importante resaltar que en los supuestos y las restricciones se deja constancia que el cliente debe proporcionar un ambiente de infraestructura para el despliegue de la solución, no se debe infringir la ley de tratamientos de datos personales y la totalidad de los costos hacen parte de un ejercicio académico que no implica un compromiso contractual entre los estudiantes y los beneficiarios. Los requisitos de aprobación son los siguientes: 1) los modelos predictivos funcionales y 2) la visualización de las cifras en tableros que permitan discriminar las diferentes características de la población objeto de estudio con el fin de facilitar la socialización de la información con la comunidad universitaria.

2.8 Aspectos éticos

De acuerdo con el código de ética publicado por la ACM, el ejercicio de la ingeniería de sistemas debe realizarse bajo una serie de principios que deben ser aplicados en la definición y ejecución durante todo el ciclo de desarrollo de software. La ACM define concretamente ocho principios, los cuales son: Publico, Cliente y Empleador, Producto, Juicio, Gestión, Profesión, Colegas, Individual. El desarrollo de este proyecto está enmarcado concretamente bajo los principios de Cliente y Empleador y Producto.

Respecto a la aplicación del principio Cliente y Empleador se dejó claridad de los siguientes aspectos:

- El cliente y/o beneficiario fue informado de la naturaleza académica de este proyecto, y así mismo tiene claro que no existe experiencia previa en la ejecución de proyectos de analítica de datos por parte de los miembros del grupo de trabajo.
- El cliente y/o beneficiario fue informado oportunamente respecto a las herramientas de desarrollo software que serán usadas durante la ejecución del proyecto. El uso de estas herramientas no compromete la adquisición de licencias pagas, ya que todas las herramientas usadas cuentan con licenciamiento *open source*, lo cual permite su uso de manera libre.
- Los artefactos generados para cada una de las fases deben ser presentados al cliente y/o beneficiario con el fin de obtener realimentación y aprobación.
- El conjunto de datos entregado por el cliente y/o beneficiario serán usados única y exclusivamente para los fines propuestos en el proyecto, adicional la información contenida debe ser anonimizada para garantizar la privacidad de los datos.

Respecto a la aplicación del principio Producto se dejó claridad de los siguientes aspectos:

- Se siguen los lineamientos por la metodología ASUM-DM, la cual es apropiada para trabajar en proyectos de analítica de datos.
- Dada la brecha de conocimiento que existe en temas de *Machine Learning*, se realiza capacitación mediante acceso a contenidos digitales en este tema, lo anterior con el fin de dar cumplimiento a los objetivos de este proyecto.

3. MARCO REFERENCIAL

De acuerdo con el marco del ciclo de transferencia tecnológica, una vez se ha identificado la problemática y se ha propuesto una solución de ingeniería, es indispensable conocer cómo se ha abordado el problema anteriormente y qué propuestas existen en entornos similares de modo que se pueda establecer un marco teórico que responde a los objetivos planteados y las necesidades de la organización.

3.1 Antecedentes y estado del arte

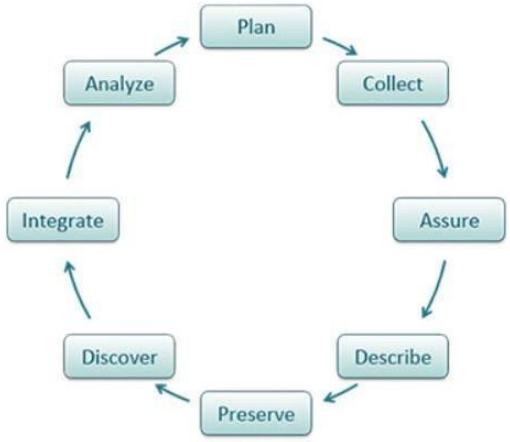
Educational Data Mining (EDM)

Los últimos avances en el sector educativo se han visto inspirados por la minería de datos en educación. Las instituciones modernas operan en un entorno competitivo complejo y el análisis del rendimiento, la provisión de una educación de alta calidad, la creación de estrategias para evaluar el rendimiento de estudiantes y profesores y la identificación de su necesidad, son los retos previos a los que se enfrentan las universidades hoy en día. [6, p. 2] La predicción del rendimiento de los estudiantes ayuda a las universidades a desarrollar y evolucionar eficazmente los planes de intervención. Proporciona excelentes beneficios para aumentar las tasas de retención, la gestión eficaz de las inscripciones, la mejora del marketing y la eficacia general de la institución. [6, p. 6]

Las instituciones almacenan una gran cantidad de información para hacer un seguimiento de los estudiantes, el profesorado y los cursos. [7, p. 1] El sistema de minería de estadísticas hace uso de KDD (*Knowledge Discovery in Datasets*), que es el procedimiento de descubrimiento de información clave a partir de un gran conjunto de datos. Este enfoque realiza una adecuada preparación, selección, limpieza e incorporación de los registros para obtener los resultados correctos. [8, p. 3470]

Toda investigación pasa por una serie de fases que determinan la gestión de la información de la tesis. Este proceso se denomina ciclo de vida de los datos [9].

Figura 6. Ciclo de vida de los datos



Fuente: Tomado de [9].

En la planificación, se establece un plan de gestión en el que se definen los datos y cómo se van a recoger y hacer accesibles a lo largo del proyecto. En la recopilación, se crea un modelo que muestra la estructura para futuros análisis. Aseguramiento se refiere a los procedimientos de calidad. En describir, se detallan exhaustivamente los estándares de metadatos adecuados. En preservar, se definen los repositorios de almacenamiento, considerando aspectos como la privacidad, la seguridad, los derechos de autor y las licencias. En integración, la información dispar se combina en un conjunto homogéneo y, finalmente, en análisis, se llevan a cabo las técnicas de experimentación definidas para extraer los resultados que permitirán confirmar o negar la hipótesis. [9]

Kishan Das Menon y Janardhan [8] proponen un procedimiento para la minería de datos en educación:

1. Los datos almacenados en la base de datos o extraídos mediante diversas herramientas de captura de datos se seleccionan como entrada para todo el proceso.
2. A continuación, los datos pasan por la etapa de preprocesamiento en la que se aplican técnicas específicas de limpieza para los valores incompletos o ausentes, o los datos ruidosos o incoherentes; todos los registros pertenecientes a determinadas divisiones se separan y se clasifican en consecuencia; la normalización es el proceso de convertir los datos al formato adecuado para su uso; la reducción hace que la información sea más escasa y densa, lo que

facilita el entrenamiento de los algoritmos para predecir y clasificar en menos tiempo.

3. La visualización es el proceso de acumulación de información y tendencias en torno a los atributos.
4. La implementación es el proceso de entrenamiento del modelo.
5. A continuación, los datos se predicen o clasifican con la ayuda del modelo antes mencionado. [8, p. 3473]

Kwok Tai Chui et al. [10] analizan los retos típicos de la analítica en un marco de *big data*:

1. Recogida: resuelve los problemas de qué tipo de datos hay que seleccionar y cuántas muestras se necesitan.
2. Análisis: se relaciona con el tratamiento de la información faltante y el preprocesamiento.
3. Privacidad: la información confidencial debe ser privada.
4. Seguridad: se trata de la retención del usuario y su confianza.
5. Algoritmos: es el proceso para obtener el mejor rendimiento en el análisis y en la implementación final. [10, p. 2]

Procedimiento de selección de los artículos

Se realizó una revisión de artículos que utilizan técnicas de aprendizaje automático para predecir y analizar el rendimiento de los estudiantes y el crecimiento poblacional. A partir de estos temas, se efectuó una búsqueda en varias bases de datos científicas: Scopus, Science Direct y Spinger. Se seleccionó un total de 47 artículos a partir de las palabras clave *"predictive"* y *"prescriptive analytics"*, y se filtró según la información del abstract. De la misma manera, se incluyeron 33 artículos con la palabra clave *"student desertion"* y *"populance variance"*. El criterio fue también la relación entre la información del resumen con el objetivo de nuestro trabajo. Se realizó una tercera búsqueda con las palabras "*machine learning AND education*", en la que se encontraron 17 artículos pertinentes. Tras analizar la relevancia de estos trabajos, finalmente, se tomaron un total de 15 artículos definitivos. En el siguiente apartado se realizará un breve resumen de las fuentes de información y se especificarán algunos de los algoritmos más eficaces para nuestro estudio.

Gothie et al. [11] trabajaron en el crecimiento y las tendencias de COVID-19 con técnicas de aprendizaje automático. El repositorio de datos fue el de la Universidad John Hopkins, que recogió 172.479 documentos sobre el tema. Para garantizar la calidad y la homogeneidad de la información, se eliminó el texto innecesario, se utilizó la lematización y la puntuación para afinar los resultados y se omitieron los términos más utilizados en la lengua inglesa (*stop words*). Para el análisis, se aplicaron los siguientes algoritmos: regresión lineal, máquina de vectores de apoyo y el modelo de previsión de series temporales Holt-Winter.

Tuli et al. [12] también predijeron el crecimiento y la tendencia de la pandemia COVID-19 con aprendizaje automático y computación en la nube. Su conjunto de datos fue Our Worl in Data, que se actualiza diariamente con los informes de la Organización Mundial de la Salud (OMS). Ellos observaron que la función Weibull inversa se ajustaba mejor al conjunto de datos del COVID-19 en comparación con las versiones iterativas de Gauss.

Panga et al. [13] analizaron el crecimiento de la obesidad infantil en Estados Unidos. Este es un problema de salud pública y se asocia a problemas de salud física y mental, como las enfermedades cardíacas y vasculares, la diabetes de tipo 2, la hipertensión y la depresión, que pueden elevar los costes de la atención sanitaria. [13, p. 1] Su estudio presenta un análisis de casi un millón de pacientes y 11 millones de registros médicos electrónicos. Incluye el control de calidad, el procesamiento y la imputación de los datos que faltan y el desarrollo de modelos de aprendizaje para la predicción de la obesidad en la primera infancia. [13 , p. 2] Todos los datos se extrajeron del repositorio Pediatric Big Data (PBD), derivado de los registros médicos del Hospital Infantil de Filadelfia. Se aplicaron múltiples comprobaciones de control de calidad para filtrar valores inverosímiles de altura, peso e índice de masa corporal (IMC).

Van Mens et al. [14] se propusieron mejorar la capacidad teórica de predicción del suicidio de los hospitales, que no ha mejorado en los últimos 50 años. Todos los factores de riesgo y de protección de la conducta suicida evaluados en el Estudio de Bienestar de Escocia se incluyen en las simulaciones. [14, p. 170]

Pallathadka et al. [7] propusieron una clasificación de los datos de rendimiento de los estudiantes, utilizando varios algoritmos de aprendizaje automático. El conjunto de datos tiene 33 atributos y 649

instancias y fue donado por la Universidad de Minho de Portugal. Según las conclusiones, los talentos e intereses de los estudiantes podrían estar relacionados con su rendimiento y, a nivel técnico, la máquina de vectores de apoyo es el algoritmo más preciso para clasificar esta variable.

Tai Chui et al. [10] llaman la atención sobre la necesidad de predecir los estudiantes universitarios en riesgo y marginales, y proponen para su estudio una máquina de soporte vectorial basada en vectores de entrenamiento reducidos. Demostraron que este enfoque podría reducir el tiempo de procesamiento sin comprometer la precisión del clasificador.

Kishan Das Menon y Janardhan [8] analizan diferentes algoritmos para extraer información sobre el proceso de aprendizaje de los estudiantes, basándose en su rendimiento anterior. El conjunto de datos consistía en detalles de 132 estudiantes de ingeniería de sexto semestre de todas las ramas del saber. Este incluyó las notas/calificaciones del estudiante desde la clase preuniversitaria hasta las disponibles del último semestre.

Zeineddine, Braendle y Farah [15] utilizaron el aprendizaje automático para aumentar la precisión de la predicción del rendimiento de los estudiantes. Consiguieron un 75,9% en total. Dabhade et al. [16] extrajeron patrones de conocimiento significativos de las bases de datos académicas. La recopilación de la información se llevó a cabo mediante una encuesta a los estudiantes y la sección académica de un instituto del sur de la India. La encuesta contó con 98 atributos y 112 preguntas. Se observó que el modelo lineal proporciona el mejor ajuste con una precisión del 83,44% y se proporcionaron pruebas de que el rendimiento pasado reciente de los estudiantes es la variable más importante para la predicción del rendimiento futuro.

Rodríguez-Hernández et al. [17] concluyeron que los modelos predictivos basados en redes neuronales artificiales muestran una buena calidad en comparación con otras metodologías predictivas y son adecuados para resolver problemas de clasificación con datos desequilibrados.

Albreiki, Zaki y Alashwal [6] hicieron una revisión sistemática de la literatura sobre la predicción del rendimiento de los estudiantes mediante el aprendizaje de máquina. Clasificaron los artículos relacionados en cuatro categorías: predicción del rendimiento de los estudiantes en riesgo; predicción del abandono de los estudiantes; predicción con datos dinámicos y estáticos y planes de recuperación

para los casos observados. A continuación, se realizará un análisis de las técnicas más relevantes para la problemática del presente proyecto.

Técnicas para la predicción

Regresión linear

La regresión lineal muestra una relación entre una variable dependiente y una o más variables independientes. Se trata de encontrar el cambio en el valor de la variable dependiente en función del valor de las variables independientes. [11, p. 3] El procedimiento es el siguiente: 1) comprobar la dirección y la correlación de los datos de entrada. 2) Ajustar la línea que pasa por los puntos, suponiendo que las variables independientes están en el eje x y las variables dependientes están en el eje y. Finalmente, 3) el modelo se evalúa en función de los siguientes parámetros: error medio absoluto, error medio cuadrático y error medio cuadrático. [8, p. 3473]

La regresión logística se utiliza para predecir la aparición de una variable dependiente categórica. El procedimiento es: 1) calcular la ocurrencia de un evento. 2) Formular el modelo, en el cual, si las variables independientes influyen en la probabilidad de la variable objetivo, se trata de una relación lineal. 3) Ajustar la regresión mediante la desestimación de máxima verosimilitud. [8, p. 3474] Para Zeineddine, Braendle et Farah [15], el método de regresión logística para predecir el rendimiento de los alumnos se utiliza normalmente para describir las asociaciones entre una serie de variables independientes que podrían clasificarse como binarias, categóricas y continuas. El nivel de precisión de la predicción mediante la regresión logística se sitúa en torno al 70% utilizando variables como las aspiraciones profesionales, el CGPA, las puntuaciones psicológicas y los intereses personales. Dabhade et al. [16] observaron que su conjunto de datos era lineal y proporcionaba el mejor *fit* con una precisión del 83,44%. El modelo que propusieron demuestra que el rendimiento pasado reciente es la variable más importante para predecir el rendimiento futuro de los estudiantes. [16, p. 5266]

Árbol de decisión

Este método de predicción es muy utilizado por su claridad y facilidad para exponer conjuntos de datos y predecir su valor. La lógica al aplicar las técnicas de árboles de decisión equivale a una serie de declaraciones IF-THEN. Un clasificador de este tipo aprende de un conjunto de puntos de datos históricos y genera la correspondiente estructura. Las características y los valores respectivos se analizan y estructuran en una topología jerárquica. Este proceso ayuda a responder a la hipótesis

recorriendo la estructura raíz-hoja. [15, p. 7] Zeineddine, Braendle y Farah [15] revelaron que varios trabajos han utilizado este método para predecir el rendimiento de los estudiantes utilizando indicadores clave como las calificaciones de los estudiantes en cursos específicos y la CGPA actual. La precisión de la predicción utilizando este método, mientras se basa en los datos previos de los estudiantes que inician un programa académico, es de alrededor del 70%, y alcanza el 90% cuando se utilizan los datos recogidos después de la incorporación al programa. [15, p. 2]

Máquina de soporte vectorial

La máquina de soporte vectorial se utiliza para problemas de clasificación y su objetivo es establecer la mejor línea para determinar un espacio en grupos, de modo que los nuevos puntos o datos puedan colocarse en la categoría correcta. [11, p. 3] Esta clasifica los puntos, dividiéndolos mediante un hiperplano de N dimensiones, donde N es el número de atributos que caracterizan un punto. Support Vector Machine (SVM) arroja los puntos de datos a un nuevo espacio de mayor dimensión en el que se vuelven linealmente separables, utilizando una función kernel específica. [15, p. 6] Zeineddine, Braendle y Farah [15] muestran que trabajos relacionados utilizaron la CGPA, las actividades extracurriculares, las pruebas psicomotoras y las evaluaciones internas para predecir el rendimiento de los estudiantes y la SVM alcanzó una precisión de alrededor del 80%. Pallathadka et al. [7] proponen que los resultados pasados de los estudiantes es uno de los criterios más significativos: "El rendimiento de los estudiantes puede anticiparse en función de sus resultados académicos anteriores. Según los resultados, los talentos e intereses de los estudiantes podrían estar relacionados con su rendimiento". [7, p. 4] La máquina de vectores de apoyo fue la técnica más precisa para este estudio.

Kwok Tai Chui et al. [10] adoptaron la SVM y propusieron un entrenamiento reducido diseñado para eliminar los vectores redundantes con el fin de bajar el tiempo de entrenamiento sin comprometer la precisión. [10, p. 2] De hecho, la evaluación del rendimiento demostró que el clasificador alcanzó una sensibilidad del 92,1-94%, una especificidad del 92-93,6% y una precisión global del 92,2-93,8% en la predicción de los estudiantes de riesgo, y una sensibilidad del 91,6-93,7%, una especificidad del 91-93,3% y una precisión global del 91,3-93,5% en la predicción de los estudiantes marginales. [10, p. 6]

Modelo Holt-Winter

El modelo de previsión de series temporales Holt-Winter es un método para modelar y proyectar el

comportamiento de un conjunto de valores en el tiempo. Es una forma de estructurar las tres facetas de las series temporales: un valor normal, una pendiente en el tiempo y una tendencia cíclica recurrente. Los valores históricos se codifican y se utilizan para predecir los valores futuros. [11, p. 3]

Función Weibull inversa

Tuli et al. [12] observaron que la función Weibull inversa se ajustaba mejor al conjunto de datos COVID-19 en comparación con las versiones iterativas de Gauss. Las predicciones del modelo gaussiano son demasiado optimistas, lo que podría llevar a un levantamiento prematuro de la cuarentena y tendría un efecto adverso en la gestión de la pandemia. Disponer de modelos que se ajusten mejor podría ayudar a planificar una estrategia basada en predicciones más precisas de los escenarios futuros [12, p. 8].

Computación en la nube

Para Tuli et al. [12] solo unas pruebas sistemáticas y planificadas pueden mitigar los efectos negativos de la propagación del COVID-19. [12, p. 8] Sistemas de vigilancia eficientes y actualizados en tiempo real pueden hacer un seguimiento adecuado de la enfermedad. Una vez que las autoridades disponen de información sobre la propagación del virus, se pueden tomar las decisiones pertinentes, como poner en cuarentena las zonas objetivo y aumentar las pruebas en las zonas adyacentes [12, p. 34]. Las instituciones gubernamentales pueden utilizar servicios en la nube para aplicar estos marcos, y, si se utilizan otros indicadores demográficos como la densidad de población, las temperaturas y la distribución por edades, se pueden hacer predicciones más fiables y precisas sobre los casos previstos [12 p. 35].

Naïve Bayes

Es una técnica de aprendizaje supervisado que se basa en el Teorema de Bayes. Asume que las características de una determinada clase son independientes, aunque sea dependiente. El procedimiento de aplicación es: 1) calcular la probabilidad de cada clase y 2) calcular la probabilidad del atributo dada la etiqueta de la clase. [8, p. 3473] Pallathadka et al. [7] aplicaron varios algoritmos de aprendizaje automático en su conjunto de datos y Naïve Bayes tuvo una precisión del 80%.

Zeineddine, Braendle y Farah [15] informaron que los trabajos que utilizaron este método consideraron predominantemente variables como las calificaciones, las becas, el CGPA, los

antecedentes de la escuela secundaria, la demografía, los datos de las redes sociales y las evaluaciones internas:

> Las investigaciones que utilizan Naïve Bayes se basan sobre todo en datos recogidos después de que los estudiantes hayan comenzado su andadura académica, con una precisión mínima del 50% y una máxima del 76%. [15, p. 3]

IDE3

Este algoritmo genera un árbol de decisión a partir del conjunto de datos. El procedimiento es el siguiente: 1) se calcula la entropía de cada atributo del conjunto de datos. 2) Se divide los atributos en subconjuntos sobre la base de dos características de mínima entropía y máxima ganancia de información. 3) Se genera el nodo de este atributo. 4) Luego, recursivamente, se crean los subconjuntos de los restantes atributos. [8, p. 3473]. Pallathadka et al. [7] informaron de una precisión del 60% del modelo utilizando esta técnica.

Red neuronal artificial

Una importante contribución de la inteligencia artificial y del área de aprendizaje automático ha sido la capacidad de construir modelos predictivos del rendimiento académico de los estudiantes mediante redes neuronales artificiales. Una red neuronal artificial (RNA) puede detectar todas las interacciones existentes entre las variables independientes. Se ha utilizado ampliamente como método en la minería de datos en educación. Las RNA también permiten el análisis de grandes volúmenes de información y la construcción de modelos predictivos independientemente de la distribución estadística de los datos [17, p. 2]. Hernández et al. demostraron que,

> Las RNA consiguen un mejor rendimiento ya que clasifican correctamente a la mayoría de los alumnos que realmente pertenecen al grupo de "alto rendimiento" (mayor recall) y, además, consiguen un mejor valor combinado entre la precisión y el recall (mayor puntuación F1).
>
> Las RNA muestran un mejor rendimiento ya que clasifican correctamente a la mayoría de los estudiantes que realmente pertenecen al "bajo rendimiento" (mayor recall) y también revelan un mejor valor combinado entre la precisión y el recall (mayor puntuación F1). [17, p. 7]

Estos autores también concluyeron que las RNA son adecuadas para resolver problemas de clasificación con datos desequilibrados y presentan ventajas sobre los resultados obtenidos al aplicar otras metodologías de predicción [17, p. 9] Zeineddine, Braendle y Farah [15] revelaron que "las variables más comunes utilizadas en la previsión del rendimiento de los estudiantes mediante redes neuronales son la actitud de los estudiantes hacia el aprendizaje, los datos de admisión, el CGPA y las calificaciones en cursos específicos" [p. 15]. Esta técnica tuvo una precisión de hasta el 98%, utilizando datos posteriores al ingreso, y tuvo una precisión de alrededor el 70% con datos anteriores al inicio de los estudios [16, p. 16].

K-NN

Es una técnica de aprendizaje supervisado que parte de la base de que existen características similares cerca unas de otras. El procedimiento es el siguiente: 1) inicializar k puntos. 2) Calcular la diferencia (distancia euclidiana) entre los datos de prueba y los datos entrenados, ordenando las diferencias. 3) De este conjunto ordenado, se seleccionan los registros y se devuelven las etiquetas. [8, p. 3473] Zeineddine, Braendle et Farah [15] descubrieron que el método K-Nearest Neighbors es rápido para predecir el rendimiento de los estudiantes en términos del nivel de aprendizaje (lento, medio, bueno y excelente). Su tasa de precisión fue superior al 60% y alcanzó el 83% cuando se utilizaron datos extraídos de las evaluaciones internas, la CGPA y las actividades extracurriculares [15, p. 3].

Automated machine learning

AutoML escoge el mejor modelo de clasificación y los correspondientes hiperparámetros para un grupo de algoritmos. Esta búsqueda concluye con un modelo de conjunto de múltiples métodos que arroja la mejor clasificación de todas las combinaciones de predicción autoprobadas. [15, p. 4] La precisión de la predicción del rendimiento de los estudiantes alcanzó el 75,9%, en general, y los autores animan a los investigadores del sector a adoptar esta técnica cuando utilicen datos previos al inicio del estudio [15, p. 9].

Preprocesamiento de datos

Para Van Mens et al. [14], el reto de predecir el comportamiento suicida es que los datos están desequilibrados, lo que significa que este tipo de observaciones son una minoría. Una técnica para hacer frente a este problema es la técnica de sobremuestreo de minorías sintéticas (Synthetic Minority Over-sampling Technique [SMOTE]). El algoritmo crea registros de entrenamiento

sintéticos para que la predicción tenga más ejemplos de comportamiento de los que aprender [14, p. 171]. Sin embargo, no se encontraron grandes diferencias entre los distintos enfoques. La regresión logística, el bosque aleatorio y el aumento del gradiente obtuvieron resultados ligeramente mejores en comparación con otros algoritmos, pero es poco probable que la diferencia sea relevante [14, p. 171].

Zeineddine, Braendle y Farah [15] se enfrentaron a tres retos principales a la hora de construir el modelo predictivo basado en su muestra: inconsistencia de los datos, desequilibrio y solapamiento. [p. 7] Se basaron en varias características para predecir el éxito de los estudiantes como las notas en los cursos clave, los exámenes, el CGPA de los últimos trimestres, las suspensiones, las advertencias, la participación en clase y los compromisos extracurriculares. Las variables representaban atributos comunes como la edad, el sexo, la etnia, el programa de estudios, la carga lectiva, la residencia en el campus, el periodo de prueba y el sistema educativo del centro. Aplicaron una cuidadosa técnica de equilibrio de datos para garantizar una mayor precisión: eligieron la técnica de sobremuestreo minoritario sintético (SMOTE) para crear puntos adicionales en el conjunto de entrenamiento con el fin de realizar un equilibrio entre las clases [15, p. 8].

Panga et al. [13] aplicaron múltiples comprobaciones de control de calidad para filtrar los valores inverosímiles de la altura, el peso y el índice de masa corporal (IMC). Se eliminaron las mediciones inconsistentes de la altura, el peso, el IMC, el peso para la altura y la circunferencia cerebral de acuerdo con las directrices de la OMS. Se identificaron los valores extremos en los signos vitales y en las pruebas de laboratorio, según los percentiles de cada variable y en la misma ventana de tiempo entre toda la población de la base de datos [13, p. 2]. A continuación, se evaluó si cada variable con información perdida terminaba completamente al azar. Las variables que no se clasificaron en este parámetro se transformaron en variables categóricas o indicadoras en las que *missing* era un indicador.

Dabhade et al. [16] corrigieron los datos añadiendo algunos atributos y generalizando las etiquetas para los datos de tipo de respuesta múltiple, aficiones e intereses. Los datos categóricos se transformaron en valores numéricos utilizando *one hot encoding* que realiza la binarización (0 o 1) de la categoría y la incluye como característica para entrenar el modelo. Gracias a los recientes avances en los sistemas de adquisición de datos y los indicadores de rendimiento del sistema, las bases de datos se estudian ahora con mayor eficacia.

En esta sección se han propuesto técnicas de minería de datos y aprendizaje automático de última generación para analizar información masiva que da lugar a un enfoque nuevo para el análisis del rendimiento de los estudiantes y el crecimiento poblacional de la población objetivo del presente trabajo.

3.2 Marco teórico

De acuerdo con Baijens, Huygh y Helms [18], el auge de las tecnologías de *big data* y las herramientas avanzadas de análisis ha generado nuevas oportunidades para generar valor a partir de los datos. Las organizaciones están asignando un número cada vez mayor de recursos a estas actividades para crear una ventaja competitiva. [18, p. 1] El análisis de datos se define como la "realización de los objetivos comerciales mediante la presentación de informes que permiten analizar tendencias, la creación de modelos predictivos para prever problemas y oportunidades futuras y la optimización de procesos para mejorar el desempeño organizacional." [18, p. 2]

En general, la gobernanza de datos se refiere a las reglas y prácticas mediante las cuales la junta directiva asegura que las estrategias sean implementadas, monitoreadas y logradas. [18, p. 3] El gobierno de las tecnologías de la información (TI) contribuye con su alineación a la estrategia de negocio para crear valor competitivo. Baijens, Huygh y Helms proponen el Modelo de Sistema Viable (VSM) [Viable System Model] como un lente teórico para establecer cómo la analítica de datos puede cumplir su propósito de crear valor empresarial a partir de la información actual y futura. Así,

> El concepto subyacente clave de VSM es la "viabilidad", que se refiere a la capacidad de un sistema para continuar cumpliendo su propósito a pesar de estar en un entorno cambiante. Como tal, un sistema viable tiene la capacidad de coevolución y adaptación dentro de un entorno dinámico. [18, p. 2]

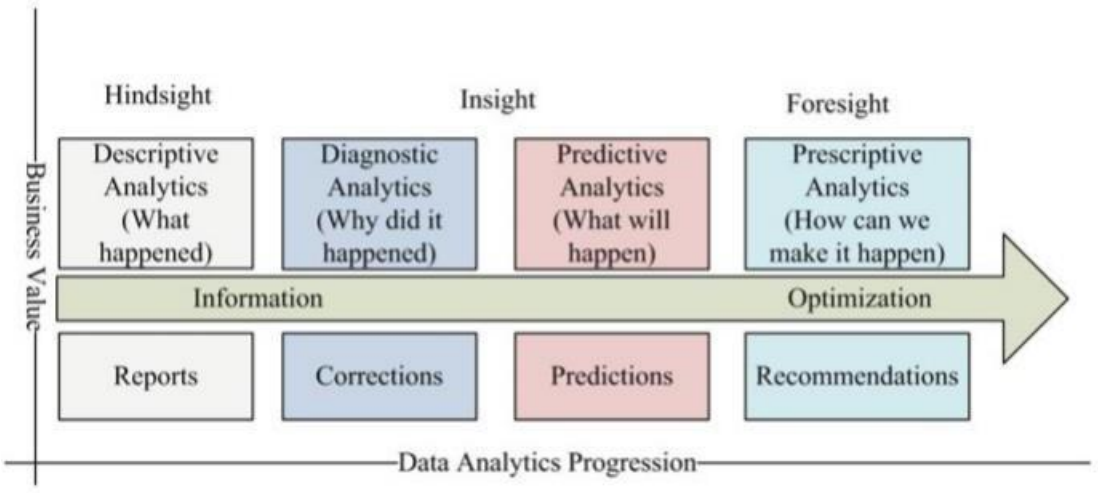
El VSM proporciona una respuesta de "organización" para hacer frente a los desafíos que surgen de operar en un entorno difícil, heterogéneo y en constante movimiento. Debido a la digitalización de los procesos, los datos son cada vez más importantes en el entorno empresarial. La creciente disponibilidad de estos se reconoce como la necesidad de tomar decisiones rápidas y basadas en hechos. [18, p. 6] Se podrían entonces enmarcar el trabajo de la Oficina de Planeación y Estadística

de la Universidad Nacional de Colombia, sede Bogotá, como un proceso de optimización de los flujos de captura, producción, procesamiento y divulgación de los datos de las diferentes dependencias con el fin de mejorar la rendición pública de cuentas y la toma de decisiones de los planes de gobierno institucionales.

La analítica de datos tiene una amplia gama de metodologías y técnicas que se clasifican, en términos generales, en descriptivos, diagnósticos, predictivos y prescriptivos. La analítica descriptiva se realiza en la etapa inicial para dar una idea razonable de la naturaleza y el patrón de los datos. Esta se concentra principalmente en el "qué" a partir de la clasificación, la agrupación y la segmentación. La siguiente etapa es concentrarse en el "por qué" ocurrió el fenómeno estudiado, enfocándose en los eventos del pasado. A este proceso se le conoce como analítica de diagnóstico.

Surge entonces la necesidad de realizar un pronóstico del futuro y se utilizan algoritmos de aprendizaje automático y técnicas de análisis estadístico para establecer tendencias de comportamientos y actividades. Sin embargo, tener una idea de los patrones de comportamiento no es suficiente para aprovechar las oportunidades comerciales, por lo tanto, la analítica prescriptiva transforma la información en conocimiento valioso para alcanzar los objetivos organizacionales de manera eficaz, determinando la mejor solución o resultado entre varias opciones según ciertos parámetros establecidos [19, p. 75].

Figura 7. Las cuatro etapas del análisis de datos



Fuente: Tomado de [19, p. 75]

En este sentido, la visualización de datos de los modelos predictivos es una herramienta esencial para la planeación estratégica institucional ya que permite reducir la incertidumbre en un entorno siempre cambiante, generar un mayor valor comercial y competitivo de la información, prever problemas y oportunidades, optimizar la gestión institucional y, en última instancia, mejorar la calidad de vida de la comunidad universitaria.

Las técnicas de *machine learning* son necesarias para mejorar la precisión de los modelos predictivos, la cual depende de la problemática a tratar y el tipo y el volumen de los datos. Muchos de los fenómenos que suceden en la Universidad Nacional no pueden ser entendidos a través de las consultas SQL a las bases de datos o los tableros interactivos de las cifras recopiladas por la OPE. Existen patrones ocultos y anomalías enterradas que estas técnicas pueden revelar. [21] Para nuestro proyecto se utilizará el aprendizaje supervisado para realizar este proceso dado que se define previamente el significado de los datos a partir de las características de los diferentes grupos poblacionales estudiados. Es importante que los científicos de datos utilicen los algoritmos correctos para mejorar el rendimiento, encontrando los datos más apropiados, limpios y precisos que reduzcan el margen de error. Así mismo, se implementará *auto machine learning* para automatizar la selección de los algoritmos y sus correspondientes parámetros de funcionamiento. Esta herramienta se basa en un procedimiento de optimización bayesiana para descubrir de manera eficiente un modelo de alta precisión para un conjunto de datos determinado. [22]

Para la metodología de trabajo se escogió ASUM-DM ya que está hace énfasis en las nuevas prácticas de la ciencia de datos como el uso de volúmenes de información muy grandes, la incorporación de análisis de texto, el modelado predictivo y la automatización de procesos. [5] Esta es una nueva versión extendida de CRISP-DM que es un modelo de estándar abierto para proyectos de minería de datos.

IBM define ASUM-DM como una guía para efectuar una implementación completa del ciclo de vida de la analítica de datos. Fue creada en 2015 para acelerar el tiempo en la generación de valor de la información y la disminución del riesgo de fracaso en los proyectos mediante la descripción de procesos estructurados que definen actividades iterativas de monitoreo y control constantes, roles, responsabilidades, plantillas y directrices generales.

Para medir la percepción de utilidad de los modelos, se utilizará la investigación cualitativa. Rolando Bolaños [23] propone un plan de ruta en su artículo "La investigación cualitativa en las ciencias de la administración: aproximaciones teóricas y metodológicas". Según este autor, las investigaciones científicas en términos generales son la vía más confiable para construir y depurar el conocimiento válido. Siguiendo el paradigma positivista, el entorno y el objeto tienen una existencia propia y están ajenos a los prejuicios y experiencias del investigador; "éste solo debe, bajo técnicas (en general, pero no exclusivamente) cuantitativas, explicar, controlar y predecir los fenómenos del entorno" [23, p. 29]. En consecuencia, la validez de los postulados es amplia de acuerdo con un margen de error predefinido y los hallazgos resultan replicables y predecibles a futuro para otros.

Por otro lado, el constructivismo establece que el conocimiento no está dado por sí, sino que debe armarse a través de "trozos" de un fenómeno. Así, "propugna la relatividad del entorno conforme el bagaje que ostenta el investigador, quien tienen interés en interpretar y comprender lo investigado en su esencia particular, sin pretender generalizar estadísticamente" [23, p. 30]

De esta forma, Bolaños presente la siguiente distinción entre dos técnicas:

> **Cuantitativa**: se fundamenta en los aspectos observables y susceptibles a cuantificar. Utiliza la metodología empírico-analítica y se sirve de la estadística para el análisis.
>
> **Cualitativa:** estudia los significados de las acciones humanas y la vida social. Se usa la metodología interpretativa como la etnografía, la fenomenología, el interaccionismo simbólico, entre otros, y su interés se centra en el descubrimiento del conocimiento. Los datos se tratan de manera explicativa. [23, p. 30]

Partiendo de estas definiciones generales, el autor propone un encuadre esquemático de un proyecto de investigación científica: formulación del problema, objetivos, marco teórico-conceptual, marco metodológico, conclusión y recomendaciones y la publicación. El elemento central de cualquier esquema es el apartado metodológico. Para el propósito del presente trabajo se utilizará la metodología de encuesta para validar el impacto de los modelos predictivos en la planeación estratégica institucional de la UNAL con el apoyo de la Oficina de Planeación y Estadística.

De acuerdo con Julio Meneses y David Rodríguez en su libro *El cuestionario y la entrevista,* el cuestionario debe atender tres requerimientos principales:

> necesidad de producir y recoger datos estructurados para tomar decisiones, gracias a la

> colaboración de las propias personas como auto-informadores, con una precisión (o error) conocida para las afirmaciones obtenidas. [23, p. 7]

El investigador no siempre puede acceder a su objeto de análisis, bien sea porque no los puede cuantificar y registrar directamente, bien porque los fenómenos resulten imposibles de ser tratados externamente como es el caso de la percepción de utilidad de un artefacto de software. En este sentido, el cuestionario se enmarca dentros de las técnicas de sistematización de autoinforme de los participantes, de modo que se estandarizan tanto las preguntas como las respuestas con el objetivo de reducir la variabilidad de la información recolectada que no se corresponda con la pregunta de investigación o hipótesis. En consecuencia, se excluyen los elementos relativos a la bifurcaciones del sentido y el significado en el discurso.

Según los autores,

> Un cuestionario es, por definición, el instrumento estandarizado que utilizamos para la recogida de datos durante el trabajo de campo de algunas investigaciones cuantitativas, fundamentalmente, las que se llevan a cabo con metodologías de encuestas. En pocas palabras, se podría decir que es la herramienta que permite al científico social plantear un conjunto de preguntas para recoger información estructurada sobre una muestra de personas, utilizando el tratamiento cuantitativo y agregado de las respuestas para describir la población a la que pertenecen o contrastar estadísticamente algunas relaciones entre variables de su interés. [23, p. 9]

De esta manera, el cuestionario es la técnica o instrumento utilizado y la metodología de encuestas es el conjunto de pasos organizados para su diseño y administración y para la recogida de los datos obtenidos. Meneses y Rodríguez [23] establecen entonces la diferenciación entre las preguntas factuales y las subjetivas. Las primeras son aquellas en las que se le pide a las personas que informes sobre hechos y acontecimientos concretos, que en principio podrán ser evaluados por el investigador. Por otra parte, las preguntas subjetivas serían aquellas en las que el ejercicio reflexivo de la persona reporta una información que no puede ser contrastada de otra manera. Es el caso de las opiniones, las creencias, los sentimientos y, en general, cualquier estado subjetivo autoinformado del que no existe ningún otro medio para acceder a él que el juicio del propio sujeto. [23, p. 12] Para este trabajo, se pretende medir la percepción de utilidad de los resultados de los modelos predictivos en la planeación estratégica institucional de la sede Bogotá de la Universidad Nacional de Colombia a partir del

autoinforme realizado por los integrantes de la Oficina de Planeación y Estadística.

De igual forma, los autores definen también las preguntas abiertas como aquellas que proporcionan el máximo grado de expresión de la respuesta verbal en un espacio de dimensiones no determinadas. Por su parte, las preguntas cerradas son aquellas en las que más allá de la escala propuesta en la encuesta, ofrecen la posibilidad al participante de escoger entre una serie de alternativas preestablecidas. [23, p. 13]

En principio, las preguntas abiertas resultan incompatibles con los fundamentos analíticos no discursivos, en otras palabras, con el tratamiento estadístico y cualitativo. Por su parte, las preguntas cerradas permiten incrementar la precisión de la respuesta, reduciendo el margen de error en la interpretación, y controlando la dispersión del sentido desde un punto de vista semántico.
De esta manera, una medida fiable es, por definición,

> Aquella que se obtiene con precisión, sin sesgos, es decir, que es consistente. En relación con el supuesto de atribución de la variabilidad, una medida fiable es aquélla en la que podemos asegurar, con un cierto nivel de confianza, que la variación observada en los datos es, de hecho, reflejo directo de la variabilidad de los fenómenos que pretendemos analizar. [...]
>
> Así, por su parte, una medida válida es aquella en la que podemos garantizar, con un cierto nivel de confianza, que estamos midiendo aquello que realmente pretendemos medir. Es decir, que es exacta y, por lo tanto, estamos dando una respuesta adecuada a la pregunta que, en último término, queremos formular mediante el uso de un cuestionario. [23, p15]

Como parte de la metodología de encuesta, los autores proponen, en primera instancia, definir la muestra, quel "no es más que un subconjunto del número total de unidades definidas como población, en referencia a la cual estableceremos siempre nuestros resultados." [23, p. 21] Es importante conocerla para comprender las limitaciones inherentes a las medidas de las variables y las conclusiones que se obtengan de ellas. Para el presente estudio, la muestra se define como los integrantes de la Oficina de Planeación y Estadística que voluntariamente quieran hacer parte del estudio.
A continuación, se plantea la necesidad de definir las preguntas y respuestas o ítems que conformarán el cuestionario. Los autores proponen cuatro escalas de medida diferentes: la escala nominal, la ordinal, la de intervalo y la de razón, también conocidas como cualitativas o no métricas, las dos primeras, y cuantitativas o métricas, las últimas. Las respuestas cerradas no ordenadas se refieren a las preguntas de

escala nominal, en la que el participante elige una de las alternativas ofrecidas. Las respuestas ordenadas implican una ordenación jerárquica entre las alternativas establecidas. Un caso particular, y de hecho habitual en la construcción de escalas, son las preguntas basadas en escalas fijas, habitualmente conocidas como tipo likert, en la que la dimensión queda ordenada en una secuencia de puntos arbitrarios de menor a mayor intensidad. Por último, las preguntas numéricas o cuantitativas son un tipo especial de pregunta abierta en la que el participante reporta un número o puntuación sobre una escala métrica de intervalo o razón. Estas representan el grado máximo de flexibilidad durante el análisis, aunque no son las más habituales en la investigación en ciencias sociales. [23. p. 16]

Para el presente proyecto se realizarán entrevistas a las personas que usen el artefacto desarrollado para medir la percepción de utilidad de este en la planeación de la UNAL. Merriam (2009), citado en Bolaños [22], plantea que el análisis cualitativo "es un proceso complejo que incluye un ir y venir entre datos concretos y abstractos, entre el razonamiento inductivo y el deductivo, y entre la descripción e interpretación de tales datos". [p. 41] Se trata de determinar la relación entre las variables establecidas de forma que fueran abordadas y explicadas entre sí y en su totalidad para comprender el fenómeno abordado de manera global. En nuestro caso se busca medir la percepción de utilidad los integrantes de la Oficina de Planeación y Estadística sobre los modelos predictivos del crecimiento poblacional de matriculados, graduados, docentes y administrativos en el marco de una investigación cualitativa con la metodología de encuesta planteada por Meneses y Rodríguez y el cuestionario como instrumento para la recopilación de los datos y su posterior análisis.

4. DESARROLLO METODOLÓGICO

De acuerdo con el ciclo de transferencia tecnológica, una vez se describe cómo se ha tratado la problemática y qué soluciones existen en entornos similares, y se ha establecido la relación de los conceptos teóricos de la Ingeniería de Sistemas con la oportunidad de mejora de los procesos organizacionales, se desarrolla, a continuación, el artefacto tecnológico orientado por los factores propios del contexto.

4.1 Fase de análisis

En la fase se realizó un análisis técnico de alto nivel (ver anexo 5) que se consolidó en el Acta de constitución del proyecto (ver anexo 6). La primera actividad del análisis fue realizar un estimado del tiempo en horas de cada uno de los bloques de la estructura de desglose de trabajo. A continuación, se identificaron y describieron los riesgos y se les asignó un responsable. En el análisis cualitativo de riesgo se calculó la probabilidad de ocurrencia de cada uno, el impacto y su correspondiente priorización. En el análisis cuantitativo correspondiente se definieron las acciones y los costos para evitar, transferir y mitigar los riesgos, o en último caso, ejecutar un plan de contingencia. Luego, se establecieron los costos del personal de acuerdo con cada riesgo.

El costo del personal del proyecto se calculó a partir del cargo, los días laborados, el salario base, las prestaciones sociales, la seguridad social, los aportes parafiscales y el número de personas. Por su parte, los costos no humanos se clasificaron a partir de cada uno de los bloques de trabajo, la descripción de los activos, la vida útil, el costo de adquisición, la depreciación mensual y anual y el tiempo del proyecto. De esta forma, se determinó que el total de costos directos es de $101,632,035 COP, total de costos no humanos es de $ 2,778,040 COP y la reserva de riesgos es de $ 48,602,250 COP, para un valor total del proyecto de $ 153,012,326 COP.

En cuanto al acta de constitución del proyecto, se definieron los entregables como los modelos predictivos del crecimiento poblacional de los grupos mencionados y los correspondientes tableros de visualización de los resultados. En cuanto a los requerimientos de alto nivel del producto, la precisión de los modelos debe ser mayor al 80% y se debe poder discriminar o filtrar por las diferentes características de la población. Se estableció, además, que se implementaría la metodología ASUM-DM como requerimiento del proyecto. Para los supuestos y las restricciones, se acordó que el beneficiario debe proporcionar un ambiente de infraestructura para el despliegue, el análisis se

realizará con datos generados por la Universidad Nacional de Colombia y deben estar anonimizados para no infringir las leyes referentes al tratamiento de datos personales y, así mismo, se hizo énfasis en que el presente proyecto es puramente académico y no implica ningún compromiso contractual con los estudiantes que pueda generar costos al beneficiario.

Al realizar el análisis técnico en la fase de inicio y planeación del proyecto, se identificaron los riesgos con su correspondientes análisis cualitativo y cuantitativo. Así, los riesgos iniciales de alto nivel son los siguientes: mala calidad y poca cantidad de los datos; hay un cambio de normatividad que afecta el acceso a estos; los modelos no generan el valor esperado al negocio, contienen sesgos o es difícil interpretar sus resultados; se realizó un levantamiento errado de los requerimientos o no se actualizaron a lo largo del proyecto; se empleó un tiempo excesivo en la construcción dado que el entrenamiento es más difícil de los esperado o no se cuenta con infraestructura necesaria para procesar grande volúmenes de datos con algoritmos complejos. En cuanto al cronograma, se aprobaron los hitos principales de la metodología seleccionada que corresponden a inicio, planeación, análisis, diseño, construcción, evaluación, testeo, despliegue, monitoreo y control, incluyendo además la fase de validación de la solución en un ambiente real, es decir, la validación dinámica, en la que se mide la percepción de utilidad del artefacto en cuestión por parte de los agentes involucrados a partir de una investigación cualitativa. Se acordó con el beneficiario y el director del proyecto de grado que se realizarán todas las fases de analítica de datos según la metodología ASUM-DM para el data set de matriculados que contiene la información histórica desde 2009 hasta el 2021. Esto con el fin de crear un prototipo inicial con las funciones mínimas para recopilar información importante sobre el desarrollo y el grado de aceptación del artefacto, centrándose en la generación de valor para la universidad.

4.2 Fase de diseño

La preparación de los datos es uno de los aspectos más importantes y, a menudo, más laboriosos de la analítica. Las actividades más importantes, dentro del proceso de análisis con datos son: seleccionar la muestra, limpiar los datos, integrarlos, derivar información clave, darle el formato adecuado y organizarlos para el modelado. En esta sección del documento se detalla cómo se realizaron estas etapas.

La Oficina de Planeación y Estadística (OPE) de la Universidad Nacional de Colombia facilitó un

archivo de Excel con el registro de los estudiantes matriculados, graduados, personal docente y personal administrativo, desde el año 2009 hasta el 2021 (ver anexo 1). Para el conjunto de datos de matriculados y graduados, los campos de la tabla (ver figura 8) corresponden al año, el semestre, el periodo (unión entre el año y el semestre, por ejemplo, 2009-2), el tipo de nivel (pregrado o posgrado), el nivel (maestría, especialización y doctorado), la facultad, el programa, el sexo de la persona, el tipo de colegio, el estrato de origen, el departamento de nacimiento, el departamento de procedencia o inscripción, la nacionalidad, la edad, el puntaje básico de matrícula, si es matriculado por primera vez, el nombre de la sede de admisión, el modo de admisión (especial o regular) y el tipo de admisión. Para el conjunto de datos de personal administrativo, los campos de la tabla corresponden al año, la sede, el sexo de la persona, la edad, el tiempo de servicio, el nivel administrativo, nivel de formación del docente.

Figura 8. Conjunto de datos matriculados

	PERIODO	NIVEL	FACULTAD	PROGRAMA	SEXO	TIPO_COL	ESTRATO_ORIG	DEP_NAC	DEP_PROC	NACIONALIDAD	EDAD_MOD	PBM_ORIG	MAT_PVEZ	SEDE_NOMBRE_ADM	TIPO_ADM	PAES	PEAMA
0	2009-1	Pregrado	Artes	Artes plásticas	Mujeres	NaN	Estrato 5	NaN	NaN	Extranjero	21.0	NaN	No	Bogotá	PAES	Mejores bachilleres	NaN
1	2009-1	Pregrado	Medicina veterinaria y de zootecnia	Medicina veterinaria	Mujeres	NaN	Estrato 4	NaN	NaN	Extranjero	21.0	NaN	No	Bogotá	Regular	NaN	NaN
2	2009-1	Pregrado	Medicina	Medicina	Hombres	NaN	Estrato 2	NaN	NaN	Extranjero	33.0	NaN	No	Bogotá	Regular	NaN	NaN
3	2009-1	Pregrado	Artes	Arquitectura	Mujeres	NaN	Estrato 3	NaN	NaN	Extranjero	33.0	NaN	No	Bogotá	Regular	NaN	NaN
4	2009-1	Pregrado	Medicina	Medicina	Hombres	NaN	Estrato 4	NaN	NaN	Extranjero	22.0	NaN	No	Bogotá	Regular	NaN	NaN

4.2.1.1 Limpieza de los datos

Sobre cada una de las variables contenida en cada uno de los conjuntos de datos se realiza un proceso de limpieza que permite depurar y establecer la calidad de los datos, la limpieza incluye la verificación de la consistencia de los datos, el manejo de datos nulos o faltantes y la selección de las variables que más relevancia tienen para análisis del crecimiento poblacional.

En primer lugar, para cada variable, se calculó la frecuencia de aparición de sus valores y se rellenaron los datos vacíos siguiendo esta distribución. Por ejemplo, se calculó la frecuencia de aparición del puntaje básico de matrícula y se generó un número aleatorio para los valores nulos de acuerdo con esta frecuencia.(Figura 9)

Figura 9. Frecuencia y función de asignación de los valores del campo puntaje básico de matrícula

```
PBM_ORIG
0.0      0.000253
1.0      0.006221
2.0      0.010971
3.0      0.012244
4.0      0.010861
           ...
96.0     0.000172
97.0     0.000305
98.0     0.000146
99.0     0.000104
100.0    0.002843
Length: 101, dtype: float64
```

```
def pbmProbabilities():
  n = np.random.choice(pbm, 1, p=f)
  n = n[0]
  return n
```

En la figura 9, el array "depNac" contiene los 32 valores que puede adquirir la variable y el array f_depNac contiene las frecuencias de aparición. Estos datos se utilizan para generar los números aleatorios. Esta misma técnica se utilizó el manejo de los datos faltantes en cada uno de los conjuntos de datos descritos previamente. Para la edad, los matriculados de 15 a 25 años, de 25 a 35, de 35 a 45, de 45 a 55 y de 55 o más (Figura 10).

Figura 10. Asignación de rangos para el campo edad

```
def asignarRango(x):
  if x < 25:
    return '15-25'
  elif 25 <= x < 35:
    return '25-35'
  elif 35 <= x < 45:
    return '35-45'
  elif 45 <= x < 55:
    return '45-55'
  elif 55 <= x:
    return '55 o más'

pregrado['EDAD'] = pregrado['EDAD_MOD'].apply(lambda x: asignarRango(x))
```

Para la variable departamento de nacimiento, se reemplazaron sus valores por las regiones delimitadas por El Órgano Colegiado de Administración y Decisión (OCAD), como se puede apreciar en la figura 11. Se definió una tabla para reemplazar los valores existentes en cada uno de los registros (ver figura 12).

Figura 11. Regiones definidas por el OCAD

Fuente. Regalías Bogotá

Figura 12. Tabla y función de asignación de las regiones OCAD

```
[ ] tabla_switch = {
        'AMAZONAS': 'CENTRO SUR',
        'ANTIOQUIA': 'EJE CAFETERO',
        'ARAUCA': 'LLANO',
        'ARCHIPIÉLAGO DE SAN ANDRÉS, PROVIDENCIA Y SANTA CATALINA': 'CARIBE',
        'ATLÁNTICO': 'CARIBE',
        'BOGOTÁ, D. C.': 'CENTRO ORIENTE',
        'BOLÍVAR': 'CARIBE',
        'BOYACÁ': 'CENTRO ORIENTE',
        'CALDAS': 'EJE CAFETERO',
        'CAQUETÁ': 'CENTRO SUR',
        'CASANARE': 'LLANO',
        'CAUCA': 'PACÍFICO',
        'CESAR': 'CARIBE',
        'CHOCÓ': 'PACÍFICO',
        'CUNDINAMARCA': 'CENTRO ORIENTE',
        'CÓRDOBA': 'CARIBE',
        'GUAINÍA': 'LLANO',
        'GUAVIARE': 'LLANO',
        'HUILA': 'CENTRO SUR',
        'LA GUAJIRA': 'CARIBE',
        'MAGDALENA': 'CARIBE',
        'META': 'LLANO',
        'NARIÑO': 'PACÍFICO',
        'NORTE DE SANTANDER': 'CENTRO ORIENTE',
        'PUTUMAYO': 'CENTRO SUR',
        'QUINDÍO': 'EJE CAFETERO',
        'RISARALDA': 'EJE CAFETERO',
        'SANTANDER': 'CENTRO ORIENTE',
        'SUCRE': 'CARIBE',
        'TOLIMA': 'CENTRO SUR',
        'VALLE DEL CAUCA': 'PACÍFICO',
        'VAUPÉS': 'LLANO',
        'VICHADA': 'LLANO'
    }
```

```
def asignaOcad(str):
    return tabla_switch.get(str, "NA")
```

```
graduadosPregrado['DEP_NAC'] = graduadosPregrado['DEP_NAC'].apply(lambda x: asignaOcad(x))
```

A continuación, se revisó la consistencia de los valores de las variables. Por ejemplo, el campo programa tenía registros mal escritos como "Ingneniería". (Figura 13)

Figura 13. Revisión de la consistencia de los valores de la variable programa

```
pregrado['PROGRAMA'].replace({"Ingneniería civil": "Ingeniería civil",
                              "Ingneniería electrónica": "Ingeniería electrónica",
                              "Ingneniería mecánica": "Ingeniería mecánica",
                              "Ingneniería química": "Ingeniería química",
                              "Ingneniería agronomica": "Ingeniería agronómica",
                              "Ingneniería industrial": "Ingeniería industrial",
                              "Ingneniería agrícola": "Ingeniería agrícola",
                              "Ingneniería mecatronica": "Ingeniería mecatrónica",
                              "Ingneniería de sistemas": "Ingeniería de sistemas",
                              "Ingneniería eléctrica": "Ingeniería eléctrica"}, inplace = True)
```

4.2.1.1 Construcción de datos

A continuación, se calculó la variable dependiente (figura 14) para cada conjunto de datos, se creó una nueva columna "Cantidad" y para cada registro se indicó el número de matriculados, graduados, docentes, administrativos de acuerdo a cada conjunto de datos.

Figura 14. Cálculo de la cantidad de matriculados por periodo y programa

```
[ ] for i, (periodop, programap, cantidadp) in enumerate(zip(pregrado['PERIODO'], pregrado['PROGRAMA'], pregrado['CantidadMatriculados'])):
      for periodogb, programagb, cantidadgb in zip(df2['PERIODO'], df2['PROGRAMA'], df2['count']):
        if (periodop == periodogb) & (programap == programagb):
          pregrado.at[i,'CantidadMatriculados'] = cantidadgb
```

```
[ ] pregrado
```

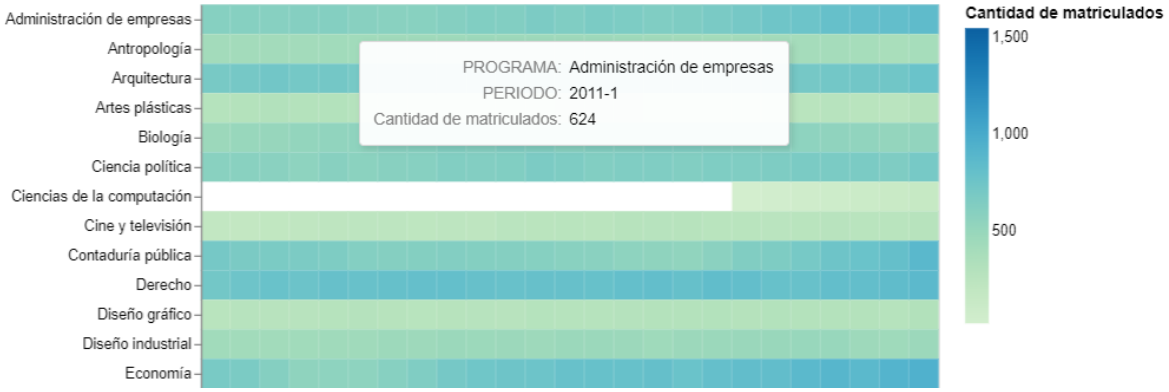
RAMA	SEXO	TIPO_COL	ESTRATO_ORIG	DEP_NAC	DEP_PROC	NACIONALIDAD	EDAD_MOD	PBM_ORIG	MAT_PVEZ	SEDE_NOMBRE_ADM	TIPO_ADM	PAES	PEAMA	PBM	EDAD	CantidadMatriculados
sticas	Mujeres	Privado	Estrato 5	CENTRO ORIENTE	LLANO	Extranjero	21.0	54.0	No	Bogotá	PAES	Mejores bachilleres	No aplica	50-59	15-25	299.0
licina naria	Mujeres	Privado	Estrato 4	LLANO	EJE CAFETERO	Extranjero	21.0	4.0	No	Bogotá	Regular	No aplica	No aplica	0-9	15-25	513.0
licina	Hombres	Privado	Estrato 2	CENTRO ORIENTE	CARIBE	Extranjero	33.0	36.0	No	Bogotá	Regular	No aplica	No aplica	30-39	25-35	1464.0
ctura	Mujeres	Privado	Estrato 3	CENTRO ORIENTE	CENTRO ORIENTE	Extranjero	33.0	42.0	No	Bogotá	Regular	No aplica	No aplica	40-49	25-35	699.0
licina	Hombres	Privado	Estrato 4	LLANO	EJE CAFETERO	Extranjero	22.0	18.0	No	Bogotá	Regular	No aplica	No aplica	10-19	15-25	1464.0

La columna "Cantidad" representa la variable objetivo para los modelos construidos, dado que se usaron algoritmos de aprendizaje automático para predecir su comportamiento con respecto a las demás variables independientes.

4.2.1.2 Análisis exploratorio

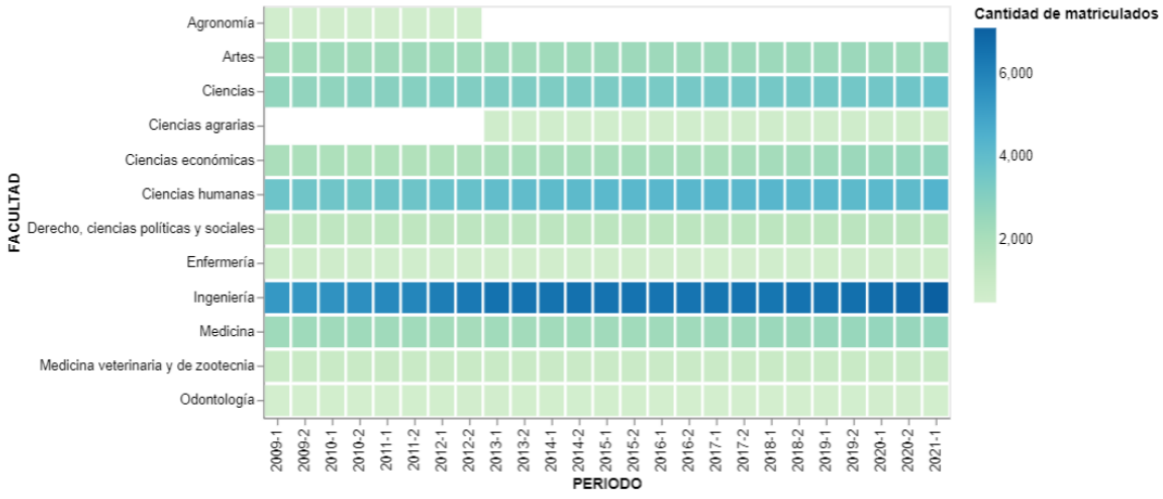
Para entender mejor el comportamiento de cada uno de los conjuntos de datos, se analizó la relación de cada variable independiente con la variable dependiente. Usando la librería de visualización Altair, se realizaron las gráficas correspondientes. En primer lugar, se realizó un agrupamiento (*groupby*) por periodo y programa. Para visualizar la tendencia de la variable objetivo en relación con las otras variables de entrada, se realizó un mapa de calor. (Figura 15).

Figura 15. Cantidad de matriculados por programa



Debido al tamaño de las gráficas, no se incluyeron en este documento la totalidad de las gráficas. Estas gráficas son útiles para ver la tendencia de los datos y compararlos con los resultados de los modelos. (Figura 16)

Figura 16. Cantidad de matriculados con respecto a la facultad



El análisis de este tipo de gráficas permite inferir información útil, por ejemplo, en la figura 16, se puede observar que las facultades con una mayor cantidad de estudiantes matriculados son medicina, ciencias humanas e ingeniería.

4.3 Fase de construcción

4.3.1 Construcción de los modelos matriculados a posgrado

La construcción de los modelos de aprendizaje automático se realiza en varias iteraciones. Por lo general, se ejecutan varios modelos con diferentes parámetros y, en varias ocasiones, se vuelve a realizar la actividad de limpieza de datos de acuerdo con las necesidades de los algoritmos. Esta sección del documento describe el proceso y selección de los modelos de predicción de los matriculados a pregrado.

La primera herramienta utilizada fue Pycaret, una biblioteca de aprendizaje automático de código abierto de Python. Es una librería integral de gestión de modelos de aprendizaje automático que acelera exponencialmente el ciclo de experimentación y lo hace más productivo. PyCaret identifica los tipos de datos contenidos para las diferentes columnas contenidas en el dataframe, en caso de existir valores faltantes, PyCaret realizará un proceso de imputación de manera automática sobre el conjunto de datos, adicionalmente, por defecto, la división de datos de entrenamiento y prueba es de 70/30.

Debido a que un conjunto de datos puede no tener la cantidad de datos suficientes con lo cual se ve afectada la precisión, PyCaret aplica N-Fold Cross Validation como estrategia para generar particiones sobre el conjunto de datos destinado para entrenamiento y prueba y construye un clasificador para cada partición y, al final, promedia los resultados de la precisión de cada partición.

Se definió un modelo de regresión de la variable objetivo ‘Cantidad’ (Figura) en cada conjunto de datos a partir de las demás variables, se realizaron 3 ciclos de selección y pruebas y no se tomaron muestras aleatorias de los datos.

Figura 17. Pycaret setup

```
from pycaret.regression import *

exp_reg101 = setup(data = data, target = 'CantidadMatriculados', fold = 3, session_id = 123,  data_split_shuffle = False, use_gpu = True)
```

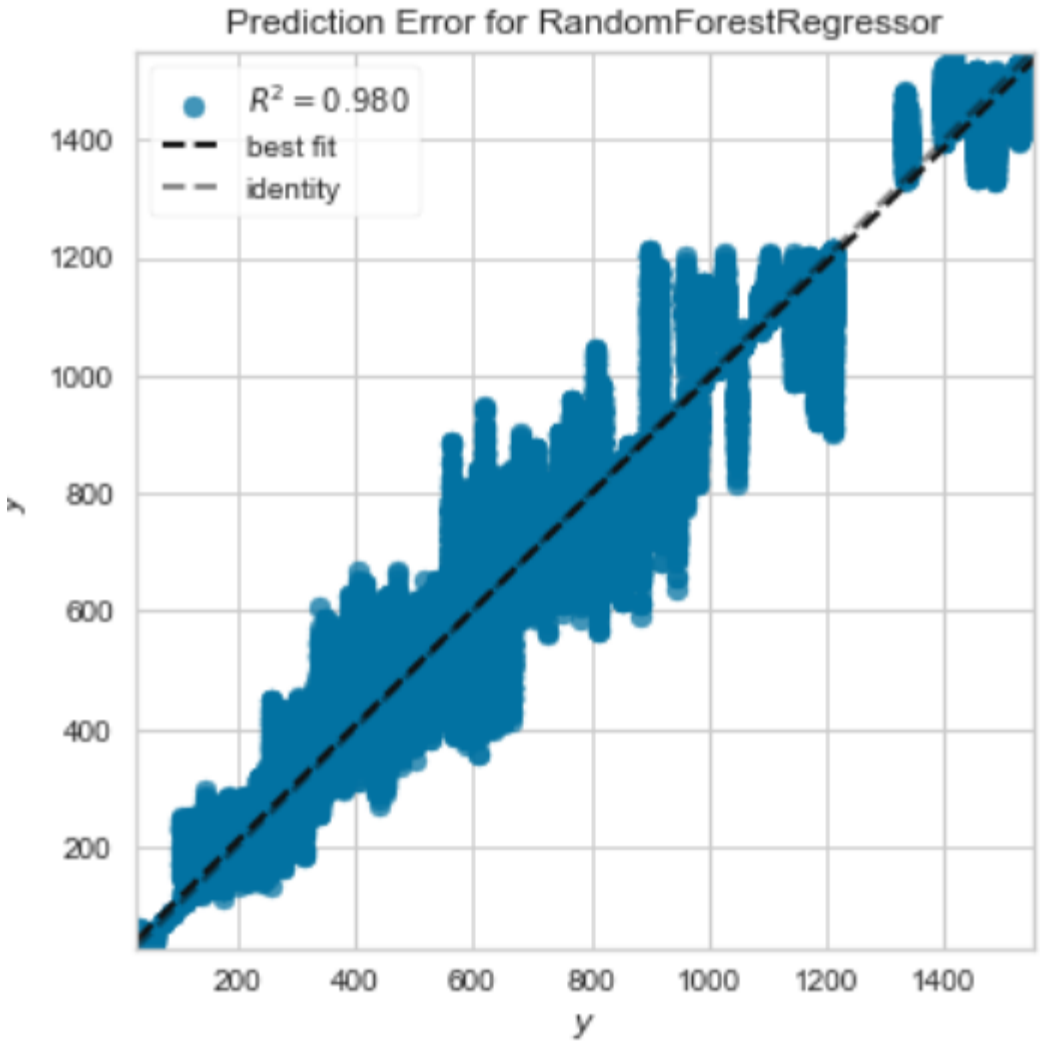
Pycaret compara varios algoritmos de regresión (figura 18) y, por defecto, los clasifica según el coeficiente de determinación (R2). Este es un número entre 0 y 1 que mide qué tan bien un modelo estadístico predice un resultado. Cuanto mejor sea un modelo para hacer predicciones, más cerca estará su R^2 de 1.

Figura 18. Comparación de algoritmos en conjunto de datos matriculados pregrado a partir del R2

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	26.9366	2020.5623	44.9503	0.9796	0.0817	0.0486	182.5967
lightgbm	Light Gradient Boosting Machine	33.1448	2069.0530	45.4868	0.9791	0.0857	0.0610	7.6900
et	Extra Trees Regressor	27.1730	2495.7226	49.9566	0.9747	0.0923	0.0491	269.2300
dt	Decision Tree Regressor	27.2075	2562.8358	50.6236	0.9741	0.0936	0.0491	10.3433
ridge	Ridge Regression	38.1898	2681.5091	51.7833	0.9729	0.0968	0.0698	1.0033
br	Bayesian Ridge	38.1882	2681.5039	51.7832	0.9729	0.0969	0.0698	15.3133
huber	Huber Regressor	37.3414	2840.9844	53.3008	0.9713	0.1034	0.0696	113.7400
par	Passive Aggressive Regressor	39.5623	3080.6858	55.5034	0.9688	0.1044	0.0731	4.5300
knn	K Neighbors Regressor	42.5200	5477.0876	74.0073	0.9446	0.1736	0.1024	3317.3000
lasso	Lasso Regression	59.0689	5550.2684	74.5000	0.9438	0.2056	0.1496	6.7200
gbr	Gradient Boosting Regressor	62.0965	5750.6323	75.8315	0.9418	0.1928	0.1454	157.1900
omp	Orthogonal Matching Pursuit	87.3481	12877.4371	113.4786	0.8697	0.2955	0.2305	1.3033
ada	AdaBoost Regressor	142.1277	29444.1308	171.5929	0.7021	0.3776	0.3426	141.4567
en	Elastic Net	212.4866	76711.0832	276.9667	0.2239	0.4755	0.4703	1.3167
llar	Lasso Least Angle Regression	234.7627	98847.0642	314.3986	-0.0000	0.5209	0.5134	1.2967
dummy	Dummy Regressor	234.7627	98847.0641	314.3986	-0.0000	0.5209	0.5134	0.2733
lr	Linear Regression	38.9353	212983.8947	299.8643	-1.1592	0.0972	0.0704	4.5800
lar	Least Angle Regression	4364168573.7943	52114675538365408870 40.0000	41751111268.8426	-52834880561397992.0000	9.7730	8436578.4617	1.3933

De acuerdo con esta búsqueda, el algoritmo más adecuado para el conjunto de datos de matriculados en pregrado es, entonces, el Random Forest Regressor o bosque aleatorio. Este ajusta una serie de árboles de decisión de clasificación en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión y controlar el sobreajuste. Pycaret ofrece una función de evaluación en la que se grafican los datos reales contra los predichos. Los resultados se pueden ver en la figura 19.

Figura 19. Error en las predicciones sobre el conjunto de datos de matriculados en pregrado



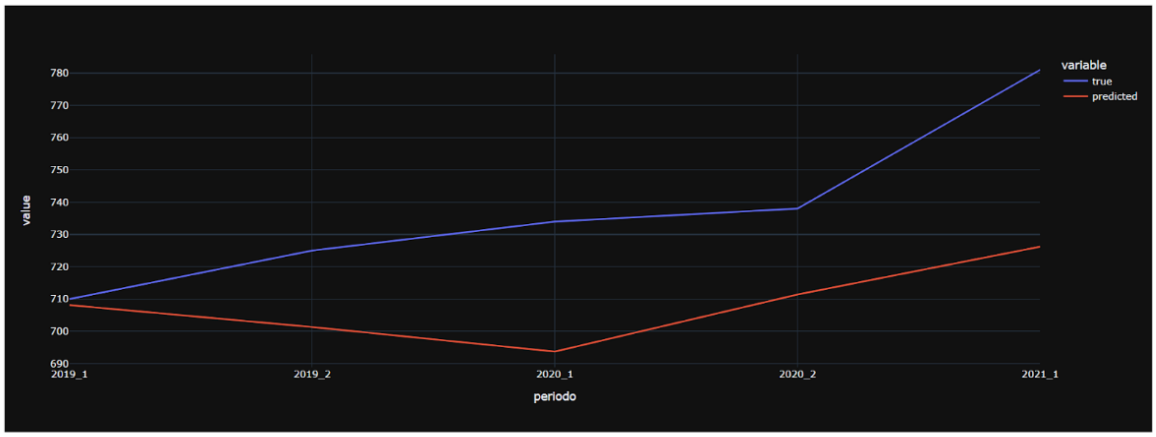
Para probar el modelo, se seleccionó el programa de arquitectura de manera aleatoria, tomando los últimos 5 semestres. Como se puede apreciar en la figura 20, el error cuadrático medio es de 26 matriculados. Se agruparon los registros por semestre y se sacó la media tanto de los valores predichos como de los valores reales. Los resultados se pueden ver en la figura 21.

Figura 20. Programa de arquitectura con Pycaret

```
from sklearn.metrics import mean_absolute_error
mse = mean_absolute_error(new_prediction['CantidadMatriculados'], new_prediction['Label'])
print(mse)
```

```
26.282893704481264
```

Figura 21. Cantidad de matriculados en programa pregrado arquitectura de 2019 a 2021



	periodo	true	predicted
0	2019_1	710.0	708.137691
1	2019_2	725.0	701.337077
2	2020_1	734.0	693.714718
3	2020_2	738.0	711.396703
4	2021_1	781.0	726.205296

A continuación, se construyó una red neuronal artificial secuencial para regresión. Para construir los vectores adecuados, se dividió el periodo en año y semestre, se borraron las columnas innecesarias, dado que no aportan información importante, y se aplicó la técnica de *one hot encoding* (figura 22), que consiste en crear una columna para cada valor categórico distinto que exista en un campo, se marca con un 1 la que concuerde con el valor del registro y se dejan las demás con 0. Teniendo en cuenta que se seleccionó el programa de arquitectura, este no cuenta con todas las columnas del dataset original, por lo que fue necesario completarlas. Así mismo, se escalaron las variables año y cantidad de

matriculados a valores entre cero y uno con standard scaler. Se dividió en variables independiente (x) y la variable objetivo (y), que es la cantidad de matriculados. Así mismo, se dividió entre datos de entrenamiento y datos de testeo con una relación de 80 a 20 por ciento y sin divisiones aleatorias.

Figura 22. Preparación de los datos para la ANN

```
pregrado = pd.get_dummies(pregrado, columns = ['FACULTAD','PROGRAMA','SEXO','TIPO_COL','ESTRATO_ORIG', 'DEP_NAC', 'DEP_PROC', 'NACIONALIDAD', 'MAT_PVEZ','TIPO_AI
pregrado.head()
```

	CantidadMatriculados	ANIO	FACULTAD_Agronomía	FACULTAD_Artes	FACULTAD_Ciencias	FACULTAD_Ciencias agrarias	FACULTAD_Ciencias económicas	FACULTAD_Ciencias humanas	FACULTAD_Derecho, ciencias políticas y sociales
0	299	2009	0	1	0	0	0	0	0
1	513	2009	0	0	0	0	0	0	0
2	1464	2009	0	0	0	0	0	0	0
3	699	2009	0	1	0	0	0	0	0
4	1464	2009	0	0	0	0	0	0	0

Finalmente, se construyó una red neuronal secuencial (figura 23) de tres capas: la primera con 128 neuronas, la segunda con 64 y la última con una de salida que corresponde al resultado de la predicción. Se realizó un entrenamiento de 50 épocas.

Figura 23. Red neuronal secuencial para regresión

```
[ ]  model = tf.keras.models.Sequential([
       tf.keras.layers.Dense(128, input_shape=(247,), activation='relu'),
       tf.keras.layers.Dense(64, activation='relu'),
       tf.keras.layers.Dense(1)
     ])
```

```
[ ]  opt = tf.keras.optimizers.Adam(0.01)
     model.compile(optimizer=opt, loss='mse', metrics=['mse'])
     r = model.fit(x_train, y_train, epochs=50, validation_data=(x_test, y_test))
```

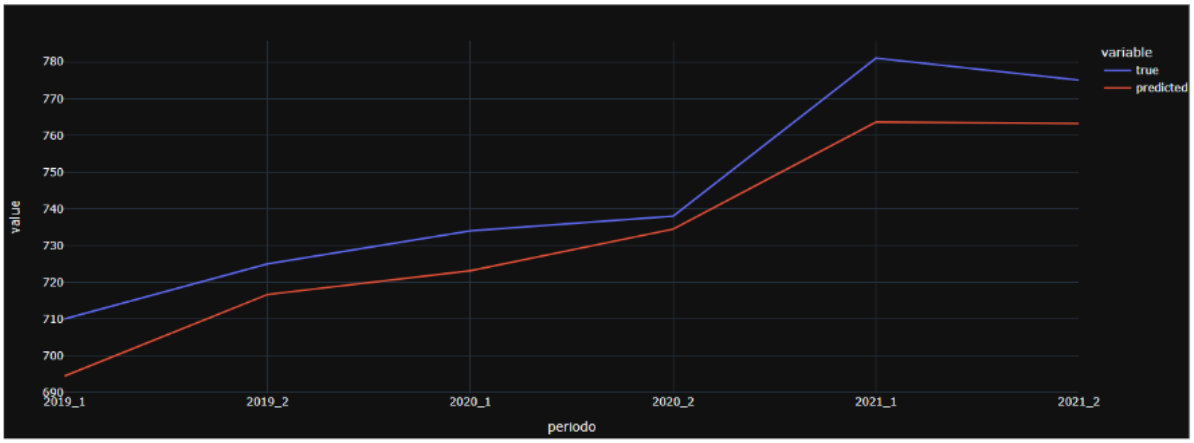
Para validar los resultados, se seleccionó el programa de arquitectura, se realizó una predicción de todos los registros y se calculó el error cuadrático medio (figura 24) con un resultado de 10 matriculados. Se tomaron los últimos cinco semestres para visualizar la comparación entre los valores reales y los predichos. Los resultados se pueden apreciar en la figura 25.

Figura 24. Error cuadrático medio en modelo ANN

```
from sklearn.metrics import mean_absolute_error
mse = mean_absolute_error(u_20212['CantidadMatriculados'], u_20212['label'])
print(mse)
```

```
11.801735827232823
```

Figura 25. Cantidad de matriculados con ANN de regresión



	periodo	true	predicted
0	2019_1	710.0	694.472992
1	2019_2	725.0	716.670550
2	2020_1	734.0	723.147718
3	2020_2	738.0	734.457509
4	2021_1	781.0	763.607524
5	2021_2	775.0	763.232092

A continuación, se utilizó la herramienta autokeras, un sistema AutoML basado en Keras, la cual es una interfaz accesible y altamente productiva para resolver problemas de aprendizaje automático, con un enfoque en el aprendizaje profundo.

Esta proporciona bloques de construcción para desarrollar y enviar soluciones de aprendizaje automático fácilmente. El objetivo de AutoKeras es hacer que el aprendizaje automático sea accesible para todos y fue desarrollado por el Data Lab de la Universidad de Texas A&M.

Para la preparación de los datos se realizaron los mismos pasos que la red neuronal de regresión. Autokeras evalúa varias estructuras o arquitecturas de redes neuronales para problemas de regresión e indica cuál es la más óptima para el conjunto de datos (ver figura 26).

Figura 26. Estructura de red propuesta por Autokeras

```
loaded_model.summary()
```

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 121)]             0

 multi_category_encoding (Mu  (None, 121)              0
 ltiCategoryEncoding)

 normalization (Normalizatio  (None, 121)              243
 n)

 dense (Dense)               (None, 32)                3904

 re_lu (ReLU)                (None, 32)                0

 dense_1 (Dense)             (None, 32)                1056

 re_lu_1 (ReLU)              (None, 32)                0

 regression_head_1 (Dense)   (None, 1)                 33

=================================================================
Total params: 5,236
Trainable params: 4,993
Non-trainable params: 243
```

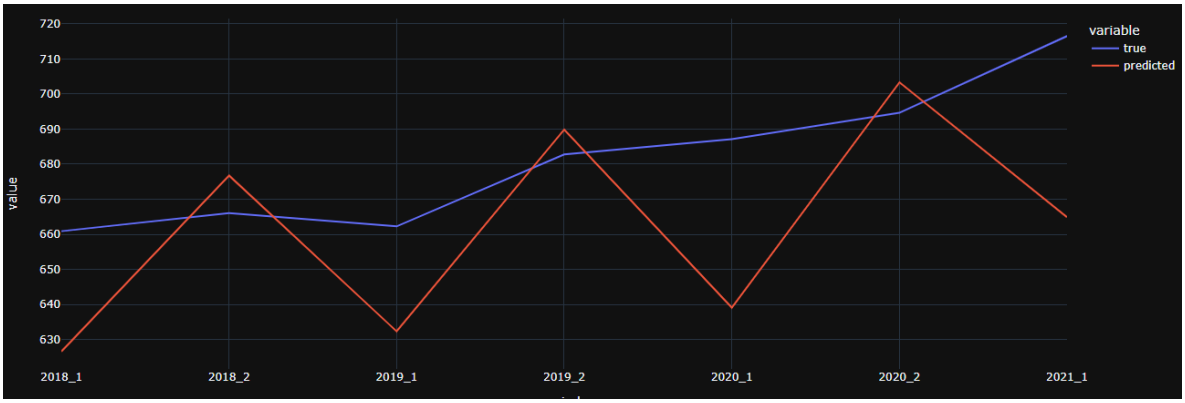
Luego, se seleccionó de igual manera el programa de arquitectura, se realizó una predicción para todos los registros (figura 27) y se sacó el error cuadrático medio con un resultado de 13 matriculados (figura 28).

Figura 27. Error cuadrático medio en modelo Autokeras

```
from sklearn.metrics import mean_absolute_error
mse = mean_absolute_error(original['CantidadMatriculados'], original['label'])
print(mse)
```

```
214.75443883003118
```

Figura 28. Cantidad de matriculados con modelo Autokeras



	periodo	true	predicted
0	2018_1	660.897689	626.629880
1	2018_2	666.029165	676.750251
2	2019_1	662.280317	632.363122
3	2019_2	682.789566	689.849025
4	2020_1	687.110544	639.081709
5	2020_2	694.636816	703.320897
6	2021_1	716.498619	664.877100

Dado que los conjuntos de datos analizados corresponden a una serie sucesiva de eventos en el tiempo, se implementó también un algoritmo de series de tiempo. Sin embargo, este enfoque presenta una dificultad: es necesario crear un modelo para cada programa, de modo que se constate la cantidad de matriculados a lo largo de todos los semestres. En efecto, se selecciona el programa de arquitectura de manera aleatoria para construir el modelo. Se utilizó una red neuronal recurrente (Long short-term memory [LSTM]). La característica principal de las redes recurrentes es que la información puede

persistir introduciendo bucles, por lo que, básicamente, pueden "recordar" estados previos y utilizar esta información para decidir cuál será el siguiente. Mientras las redes recurrentes estándar pueden modelar dependencias a corto plazo (es decir, relaciones cercanas en la serie cronológica), las LSTM pueden aprender dependencias largas, por lo que se podría decir que tienen una "memoria" a largo plazo. Por último, se dividió entre las variables independientes (input) y la variable independiente (target).

Para la serie de tiempo, se construyó una ventana deslizante (figura 29), es decir, se utilizaron los valores de registros anteriores para predecir el siguiente en cada una de las columnas. El tamaño de la ventana escogido fue 7. Así, para cada campo y cada registro, se extrajeron siete columnas más que corresponden a los siete valores anteriores, de forma que se pudiera predecir el valor actual. De esta forma, se crearon los datos de entrenamiento con dos tercios del dataset y el resto como datos de testeo.

Figura 29. Construcción de la ventana deslizante

```
T = 7
D = input_data.shape[1]
N = len(input_data) - T
```

```
Ntrain = len(input_data) * 2 // 3
```

```
X_train = np.zeros((Ntrain,T,D))
Y_train = np.zeros(Ntrain)
for t in range(Ntrain):
   X_train[t, :, :] = input_data[t:t+T]
   Y_train[t] = target[t+T]
```

```
X_test = np.zeros((N - Ntrain,T,D))
Y_test = np.zeros(N - Ntrain)
for u in range(N - Ntrain):
  t = u + Ntrain
  X_test[u, :, :] = input_data[t:t+T]
  Y_test[u] = target[t+T]
```

Luego, se construyó una red neuronal recurrente con una capa input, una capa de 50 neuronas LSTM (figura 30) y una capa de salida con un único valor. La métrica de pérdida fue el error cuadrático

medio.

Figura 30. LSTM

```
[ ]  i = Input(shape=(T, D))
     x = LSTM(50)(i)
     x = Dense(1)(x)
     model = Model(i, x)
     model.compile(loss='mse', optimizer=Adam(lr=0.01))
```

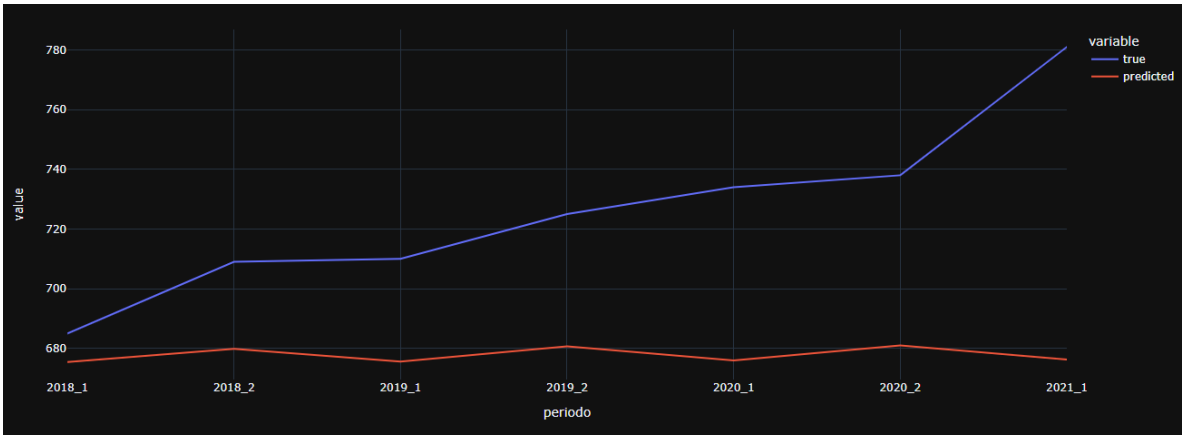
Se realizaron 100 épocas de entrenamiento. El error cuadrático medio de la predicción (figura 31) fue de 4 matriculados. Para evaluar el modelo se extrajeron los registros de los últimos semestres y se les sacó el promedio a los valores reales y las predicciones. Los resultados se pueden apreciar en la figura 32.

Figura 31. Error cuadrático medio en modelo Time Series

```
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(ArquitecturaOriginal['CantidadMatriculados'], ArquitecturaOriginal['label'])
print(mae)
```

```
14.834229828510788
```

Figura 32. Cantidad de matriculados a arquitectura LSTM



	periodo	true	predicted
0	2018_1	685.0	675.435533
1	2018_2	709.0	679.879786
2	2019_1	710.0	675.614509
3	2019_2	725.0	680.689890
4	2020_1	734.0	675.955263
5	2020_2	738.0	680.961444
6	2021_1	781.0	676.253336

Teniendo en cuenta que es necesario construir un modelo para cada programa en las series de tiempo, se consideró como inviable e insostenible en el tiempo este enfoque pues habría que entrenar de nuevo la red neuronal para cada programa todos los semestres. Además, los resultados no son necesariamente mejores a las demás técnicas. A continuación, se pasa al proceso de evaluación de los modelos anteriores con datos que corresponden al semestre de 2021-2, los cuales son enteramente nuevos para estos. Este proceso de limpieza y construcción se replicó para todo los conjuntos de datos de matriculados, graduados, docentes y administrativos, seleccionando finalmente los modelos con la mejor predicción. La especificación de cada uno de estos de ellos queda especificado en los reportes que se le entregaron al cliente (ver anexo 8).

5. RESULTADOS

La evaluación de los modelos se realiza utilizando los criterios de éxito del contexto del negocio planteados por el cliente. Esto se realiza para entender e interpretar los resultados de acuerdo a los objetivos del proyecto y la línea de conocimiento base y así ver su valor de utilidad en un entorno real. Es importante revisar los procesos que se llevaron a cabo para analizar qué es necesario, qué fue ejecutado exitosamente y qué se puede mejorar e identificar las fallas, los pasos en falso y los caminos alternativos para explorar. Si los procesos y los modelos satisfacen los objetivos planteados, se puede seguir a la fase de despliegue.

Validación estática

En el ciclo de transferencia tecnológica de ingeniería, una vez desarrollado el artefacto, se procede a la validación estática del mismo en un ambiente de prueba previo al de producción. En el presente proyecto de análisis, se pretende evaluar la calidad de las predicciones de los modelos en comparación con los datos reales.

Para el conjunto inicial de matriculados, se cuenta con los registros desde 2009 hasta el primer periodo de 2021. El cliente facilitó los datos del segundo periodo de 2021 para evaluar la precisión de los modelos construidos. A este conjunto de datos se le realizó el mismo trabajo de limpieza que al dataset original, es decir, se dividió entre pregrado y posgrado, se rellenaron los datos faltantes, se revisó la consistencia y se creó la variable objetivo "Cantidad de matriculados". Teniendo en cuenta que para el periodo 2021-2 se pueden haber creado programas nuevos, se filtró por aquellos que son conocidos por el modelo. Finalmente, se concatenaron los dos datasets. A continuación, para matriculados en posgrado, se utilizaron los dos mejores modelos: el *Random Forest Regresso*r, obtenido a partir de Pycaret y la red neuronal artificial de regresión. Como se puede ver en la figura 33, para el primer algoritmo, se seleccionaron al azar las predicciones de los últimos cinco años de las maestría de urbanismo. El error absoluto medio fue de 3 matriculados.

Figura 33. Matriculados en posgrado con Random Forest Regressor

```
[ ]  programa = new_prediction[new_prediction['NIVEL'] == 'Maestría']

[ ]  programa = programa[programa['PROGRAMA'] == 'Urbanismo']

[ ]  programa = programa[programa['PERIODO'] > '2018-01-01']
```

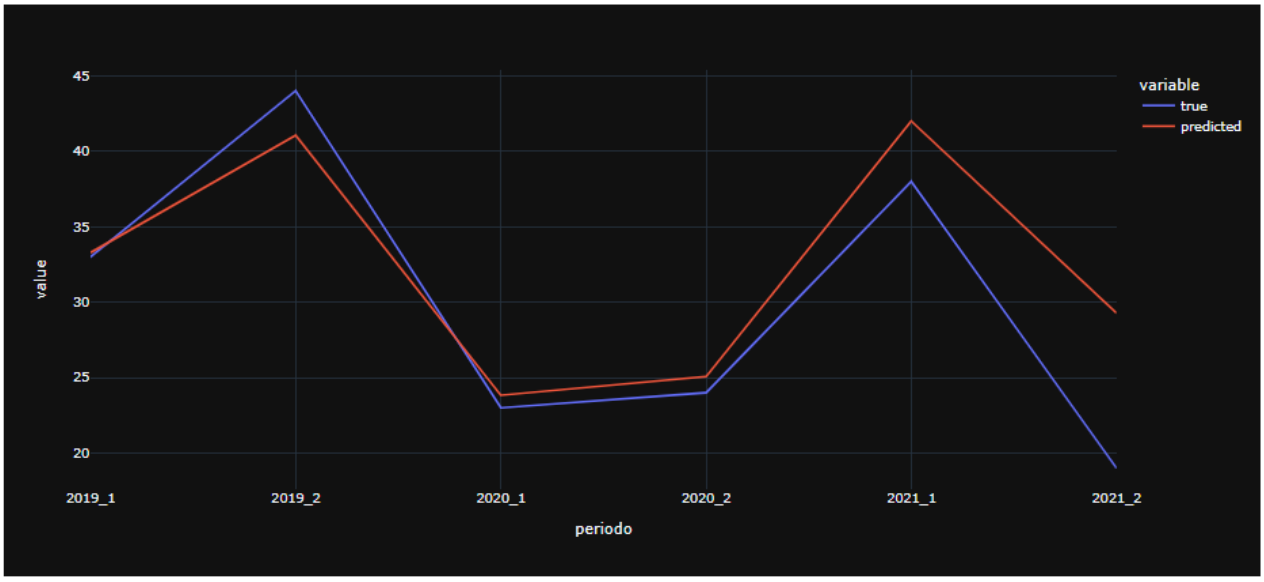
```
from sklearn.metrics import mean_absolute_error
mse = mean_absolute_error(programa['CantidadMatriculados'], programa['Label'])
print(mse)
```

```
2.4805777537382236
```

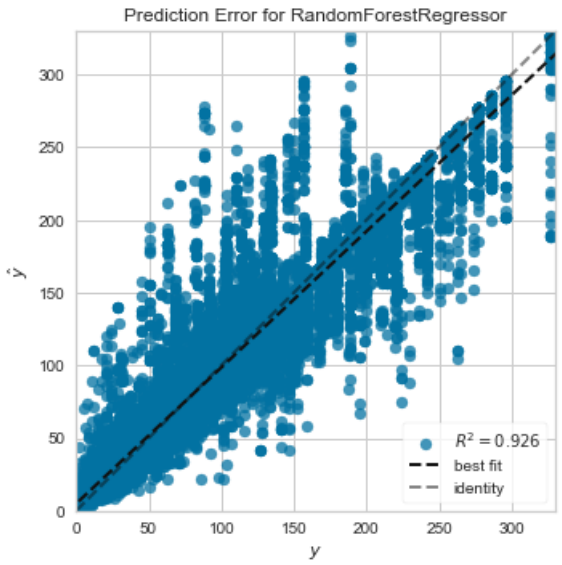
Así mismo, se construyó la gráfica comparativa entre los valores predichos y los valores reales como se puede apreciar en la figura 34.

Figura 34. Valores reales versus valores predichos con Random Forest Regressor

```
a_20191 = programa[programa['PERIODO'] == '2019-1']
t = a_20191['CantidadMatriculados'].mean()
p = a_20191['Label'].mean()
a_20197 = programa[programa['PERIODO'] == '2019-7']
t2 = a_20197['CantidadMatriculados'].mean()
p2 = a_20197['Label'].mean()
a_20201 = programa[programa['PERIODO'] == '2020-1']
t3 = a_20201['CantidadMatriculados'].mean()
p3 = a_20201['Label'].mean()
a_20207 = programa[programa['PERIODO'] == '2020-7']
t4 = a_20207['CantidadMatriculados'].mean()
p4 = a_20207['Label'].mean()
a_20211 = programa[programa['PERIODO'] == '2021-1']
t5 = a_20211['CantidadMatriculados'].mean()
p5 = a_20211['Label'].mean()
a_20212 = programa[programa['PERIODO'] == '2021-7']
t6 = a_20212['CantidadMatriculados'].mean()
p6 = a_20212['Label'].mean()
```



	periodo	true	predicted
0	2019_1	33.0	33.284794
1	2019_2	44.0	41.046737
2	2020_1	23.0	23.834595
3	2020_2	24.0	25.072150
4	2021_1	38.0	41.994884
5	2021_2	19.0	29.284737



Se realizó el mismo procedimiento para la red neuronal artificial y se seleccionaron las predicciones de los últimos cinco años de la maestría de arquitectura. El error absoluto medio fue de 7 matriculados. Los resultados se pueden apreciar en la figura 35.

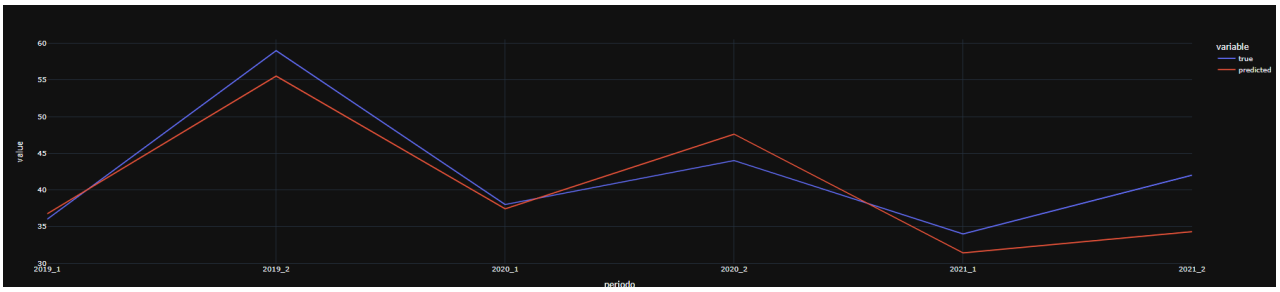
Figura 35. Matriculados en posgrado con ANN

```
u_20191 = postgrado[postgrado['NIVEL_Maestría'] == 1]
u_20191 = u_20191[u_20191['PROGRAMA_Arquitectura'] == 1]
u_20191 = u_20191[u_20191['ANIO'] == 2019]
u_20191 = u_20191[u_20191['SEMESTRE_1'] == 1]
t = u_20191['CantidadMatriculados'].mean()
p = u_20191['label'].mean()

u_20192 = postgrado[postgrado['NIVEL_Maestría'] == 1]
u_20192 = u_20192[u_20192['PROGRAMA_Arquitectura'] == 1]
u_20192 = u_20192[u_20192['ANIO'] == 2019]
u_20192 = u_20192[u_20192['SEMESTRE_2'] == 1]
t2 = u_20192['CantidadMatriculados'].mean()
p2 = u_20192['label'].mean()

u_20201 = postgrado[postgrado['NIVEL_Maestría'] == 1]
u_20201 = u_20201[u_20201['PROGRAMA_Arquitectura'] == 1]
u_20201 = u_20201[u_20201['ANIO'] == 2020]
u_20201 = u_20201[u_20201['SEMESTRE_1'] == 1]
t3 = u_20201['CantidadMatriculados'].mean()
p3 = u_20201['label'].mean()

u_20202 = postgrado[postgrado['NIVEL_Maestría'] == 1]
u_20202 = u_20202[u_20202['PROGRAMA_Arquitectura'] == 1]
u_20202 = u_20202[u_20202['ANIO'] == 2020]
u_20202 = u_20202[u_20202['SEMESTRE_2'] == 1]
t4 = u_20202['CantidadMatriculados'].mean()
p4 = u_20202['label'].mean()

u_20211 = postgrado[postgrado['NIVEL_Maestría'] == 1]
u_20211 = u_20211[u_20211['PROGRAMA_Arquitectura'] == 1]
u_20211 = u_20211[u_20211['ANIO'] == 2021]
u_20211 = u_20211[u_20211['SEMESTRE_1'] == 1]
t5 = u_20211['CantidadMatriculados'].mean()
p5 = u_20211['label'].mean()


u_20212 = postgrado[postgrado['NIVEL_Maestría'] == 1]
u_20212 = u_20212[u_20212['PROGRAMA_Arquitectura'] == 1]
u_20212 = u_20212[u_20212['ANIO'] == 2021]
u_20212 = u_20212[u_20212['SEMESTRE_2'] == 1]
t6 = u_20212['CantidadMatriculados'].mean()
p6 = u_20212['label'].mean()
```

```
[ ] from sklearn.metrics import mean_absolute_error
    mse = mean_absolute_error(u_20212['CantidadMatriculados'], u_20212['label'])
    print(mse)
```

```
7.697775929989371
```

Figura 36. Valores reales versus valores predichos con ANN



	periodo	true	predicted
0	2019_1	36.0	36.721805
1	2019_2	59.0	55.530293
2	2020_1	38.0	37.418878
3	2020_2	44.0	47.598866
4	2021_1	34.0	31.404795
5	2021_2	42.0	34.302224

Para el conjunto de matriculados en pregrado se aplicó el mismo proceso. Se seleccionaron los modelos de Random Forest Regressor y la red neuronal de regresión. Se tomaron los últimos cinco años del programa de medicina para evaluar la predicción. El error absoluto medio fue de 24 matriculados. Los resultados se pueden ver en la gráfica 37.

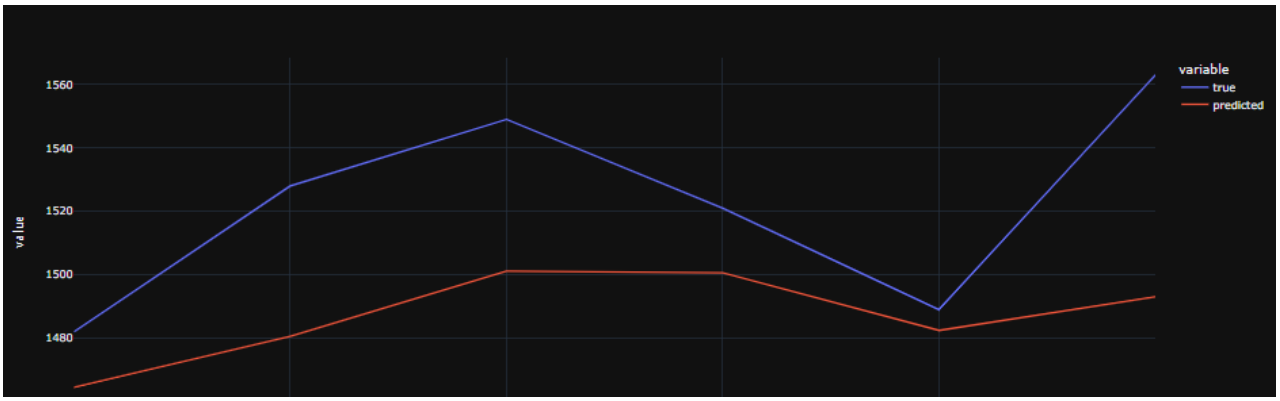
Figura 37. Matriculados en pregrado con Random Forest Regressor

```
[ ] Arquitectura = new_prediction[new_prediction['PROGRAMA'] == 'Medicina']

[ ] Arquitectura = Arquitectura[Arquitectura['PERIODO'] > '2009-01-01']
```

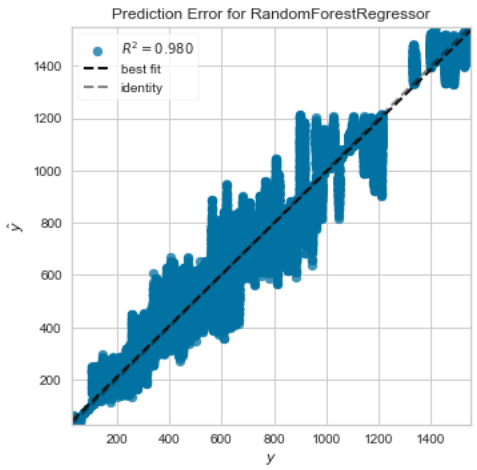
```
from sklearn.metrics import mean_absolute_error
mse = mean_absolute_error(Arquitectura['CantidadMatriculados'], Arquitectura['Label'])
print(mse)
```

```
24.410021071860054
```

Figura 38. Valores reales versus valores predichos con Random Forest Regressor



	periodo	true	predicted
0	2019_1	1482.0	1464.468960
1	2019_2	1528.0	1480.534179
2	2020_1	1549.0	1501.125065
3	2020_2	1521.0	1500.557197
4	2021_1	1489.0	1482.418754
5	2021_2	1563.0	1493.000000



En el caso de la red neuronal se seleccionaron los últimos semestres del programa de arquitectura. El error absoluto medio fue de 11 matriculados.

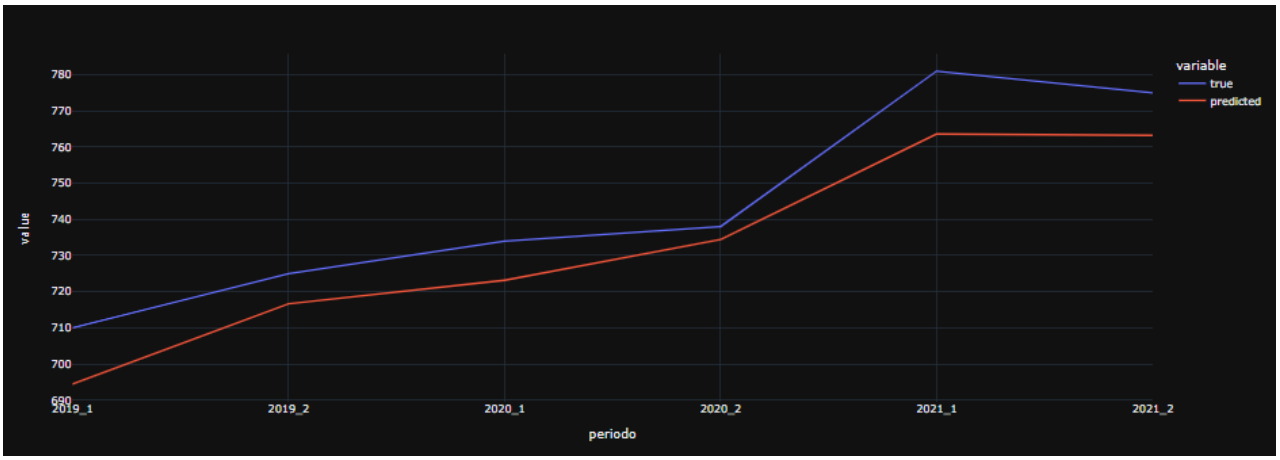
Figura 39. Matriculados en pregrado con ANN

```
u_20191 = pregrado[pregrado['PROGRAMA_Arquitectura'] == 1]
u_20191 = u_20191[u_20191['ANIO'] == 2019]
u_20191 = u_20191[u_20191['SEMESTRE_1'] == 1]
t = u_20191['CantidadMatriculados'].mean()
p = u_20191['label'].mean()

u_20192 = pregrado[pregrado['PROGRAMA_Arquitectura'] == 1]
u_20192 = u_20192[u_20192['ANIO'] == 2019]
u_20192 = u_20192[u_20192['SEMESTRE_2'] == 1]
t2 = u_20192['CantidadMatriculados'].mean()
p2 = u_20192['label'].mean()

u_20201 = pregrado[pregrado['PROGRAMA_Arquitectura'] == 1]
u_20201 = u_20201[u_20201['ANIO'] == 2020]
u_20201 = u_20201[u_20201['SEMESTRE_1'] == 1]
t3 = u_20201['CantidadMatriculados'].mean()
p3 = u_20201['label'].mean()

u_20202 = pregrado[pregrado['PROGRAMA_Arquitectura'] == 1]
u_20202 = u_20202[u_20202['ANIO'] == 2020]
u_20202 = u_20202[u_20202['SEMESTRE_2'] == 1]
t4 = u_20202['CantidadMatriculados'].mean()
p4 = u_20202['label'].mean()


u_20211 = pregrado[pregrado['PROGRAMA_Arquitectura'] == 1]
u_20211 = u_20211[u_20211['ANIO'] == 2021]
u_20211 = u_20211[u_20211['SEMESTRE_1'] == 1]
t5 = u_20211['CantidadMatriculados'].mean()
p5 = u_20211['label'].mean()

u_20212 = pregrado[pregrado['PROGRAMA_Arquitectura'] == 1]
u_20212 = u_20212[u_20212['ANIO'] == 2021]
u_20212 = u_20212[u_20212['SEMESTRE_2'] == 1]
t6 = u_20212['CantidadMatriculados'].mean()
p6 = u_20212['label'].mean()
```

```
from sklearn.metrics import mean_absolute_error
mse = mean_absolute_error(u_20212['CantidadMatriculados'], u_20212['label'])
print(mse)
```

```
11.801735827232823
```

Figura 40. Valores reales versus valores predichos con ANN



	periodo	true	predicted
0	2019_1	710.0	694.472992
1	2019_2	725.0	716.670550
2	2020_1	734.0	723.147718
3	2020_2	738.0	734.457509
4	2021_1	781.0	763.607524
5	2021_2	775.0	763.232092

Para efectuar la validación estática del dataset de graduados, se dividió asimismo entre pregrado y posgrado, se entrenó el modelo con los datos hasta 2021-1 y se realizó la evaluación con los datos para el 2021-2. De igual forma, se realizó la limpieza de datos, se reemplazaron los valores nulos, se revisó la integridad de los datos y se construyó la variable objetivo. Teniendo en cuenta que pueden haber programas nuevos cada semestre, se filtró con aquellos conocidos hasta el entrenamiento. Los modelos seleccionados fueron también el *random forest regressor* y la red neuronal de regresión. Se seleccionaron los últimos semestres del programa de la maestría en urbanismo para evaluar la predicción. El error absoluto medio fue de 1 graduado y los resultados se pueden ver en las gráficas a continuación.

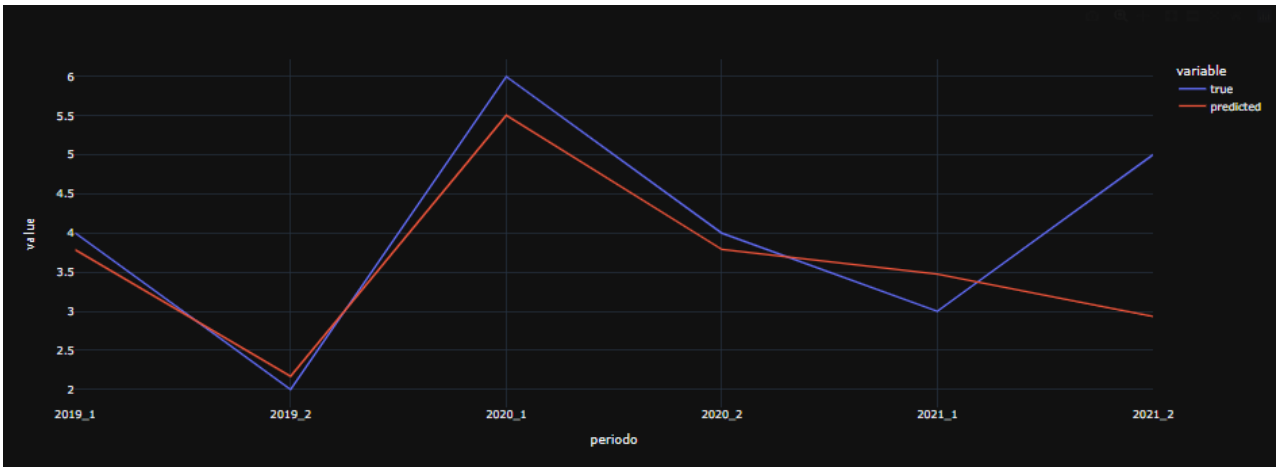
Figura 41. Graduados en posgrado con Random Forest Regressor

```
programa = new_prediction[new_prediction['NIVEL'] == 'Maestría']

programa = programa[programa['PROGRAMA'] == 'Urbanismo']

programa = programa[programa['PERIODO'] > '2018-01-01']
```

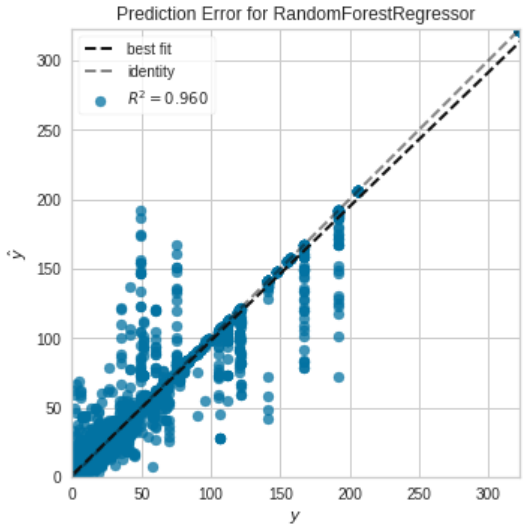
```
from sklearn.metrics import mean_absolute_error
mse = mean_absolute_error(programa['CantidadMatriculados'], programa['Label'])
print(mse)
```

```
0.7385185185185186
```

Figura 42. Valores reales versus valores predichos con Random Forest Regressor



	periodo	true	predicted
0	2019_1	4.0	3.787500
1	2019_2	2.0	2.170000
2	2020_1	6.0	5.505000
3	2020_2	4.0	3.790000
4	2021_1	3.0	3.476667
5	2021_2	5.0	2.936000



Para los graduados en pregrado, se seleccionó el programa de Economía y se evaluó las predicciones para el segundo semestre de 2021, con un error absoluto de 4 matriculados.

Figura 43. Graduados en pregrado con Random Forest Regressor

```
programa = new_prediction[new_prediction['NIVEL'] == 'Pregrado']
```

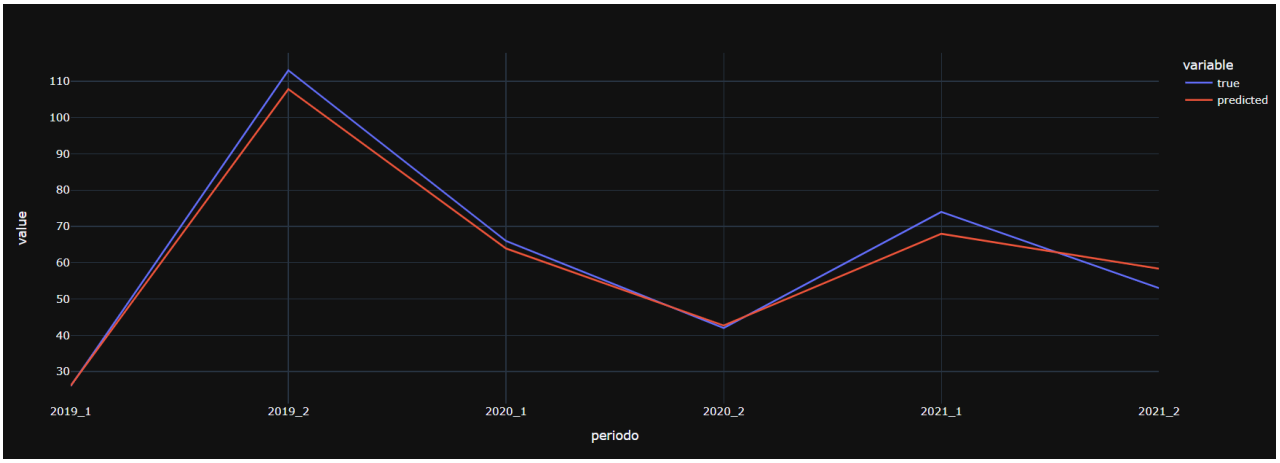
```
programa = programa[programa['PROGRAMA'] == 'Economía']
```

```
programa = programa[programa['PERIODO'] > '2018-01-01']
```

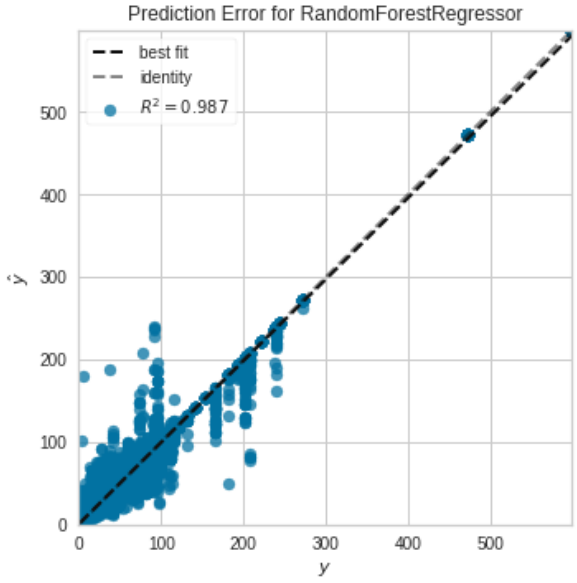
```
[ ] from sklearn.metrics import mean_absolute_error
    mse = mean_absolute_error(programa['CantidadMatriculados'], programa['Label'])
    print(mse)
```

```
4.300826912780437
```

```
a_20191 = programa[programa['PERIODO'] == '2019-1']
t = a_20191['CantidadMatriculados'].mean()
p = a_20191['Label'].mean()
a_20197 = programa[programa['PERIODO'] == '2019-7']
t2 = a_20197['CantidadMatriculados'].mean()
p2 = a_20197['Label'].mean()
a_20201 = programa[programa['PERIODO'] == '2020-1']
t3 = a_20201['CantidadMatriculados'].mean()
p3 = a_20201['Label'].mean()
a_20207 = programa[programa['PERIODO'] == '2020-7']
t4 = a_20207['CantidadMatriculados'].mean()
p4 = a_20207['Label'].mean()
a_20211 = programa[programa['PERIODO'] == '2021-1']
t5 = a_20211['CantidadMatriculados'].mean()
p5 = a_20211['Label'].mean()
a_20212 = programa[programa['PERIODO'] == '2021-7']
t6 = a_20212['CantidadMatriculados'].mean()
p6 = a_20212['Label'].mean()
```



	periodo	true	predicted
0	2019_1	26.0	26.093077
1	2019_2	113.0	107.828568
2	2020_1	66.0	63.920580
3	2020_2	42.0	42.680179
4	2021_1	74.0	67.960675
5	2021_2	53.0	58.332757

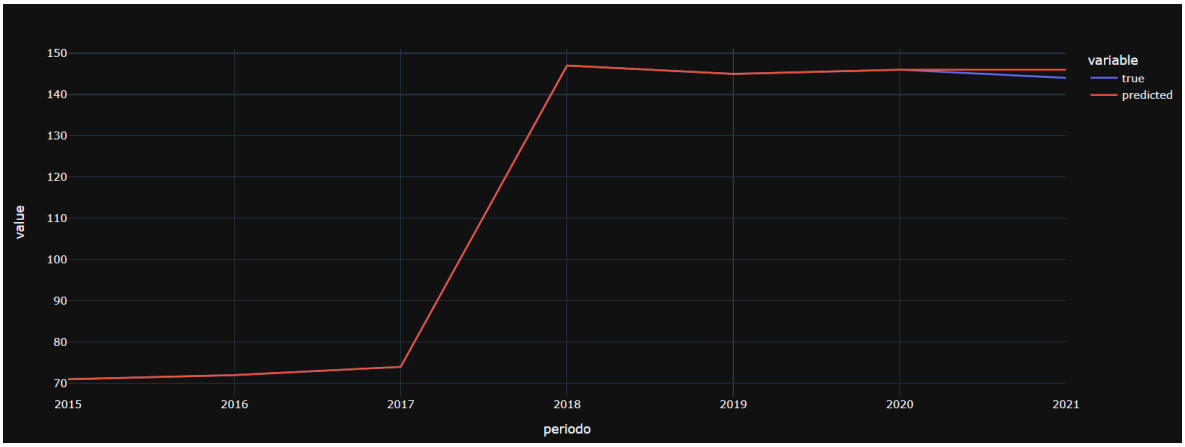


Para los dataset de docentes y administrativos, la construcción y evaluación de los modelos se llevó a cabo siguiendo el mismo patrón que la evaluación estática de matriculados y graduados, es decir, el entrenamiento del modelo se construyó con los datos hasta 2020 y se evaluó con los datos del 2021. Se escogió el algoritmo *Random Forest Regressor* para ambos dado la calidad de su precisión.

Figura 44. Cantidad de docentes de 2018 a 2021

```
df = predictionsAll[predictionsAll['UNIDAD'] == 'Departamento de Física']
```

```
df2015 = df[df['YEAR'] == 2015]
t2015 = df2015['CantidadDocentes'].mean()
p2015 = df2015['Label'].mean()

df2016 = df[df['YEAR'] == 2016]
t2016 = df2016['CantidadDocentes'].mean()
p2016 = df2016['Label'].mean()

df2017 = df[df['YEAR'] == 2017]
t2017 = df2017['CantidadDocentes'].mean()
p2017 = df2017['Label'].mean()

df2018 = df[df['YEAR'] == 2018]
t2018 = df2018['CantidadDocentes'].mean()
p2018 = df2018['Label'].mean()

df2019 = df[df['YEAR'] == 2019]
t2019 = df2019['CantidadDocentes'].mean()
p2019 = df2019['Label'].mean()

df2020 = df[df['YEAR'] == 2020]
t2020 = df2020['CantidadDocentes'].mean()
p2020 = df2020['Label'].mean()

df2021 = df[df['YEAR'] == 2021]
t2021 = df2021['CantidadDocentes'].mean()
p2021 = df2021['Label'].mean()
```



	periodo	true	predicted
0	2015	71.0	71.0
1	2016	72.0	72.0
2	2017	74.0	74.0
3	2018	147.0	147.0
4	2019	145.0	145.0
5	2020	146.0	146.0
6	2021	144.0	146.0

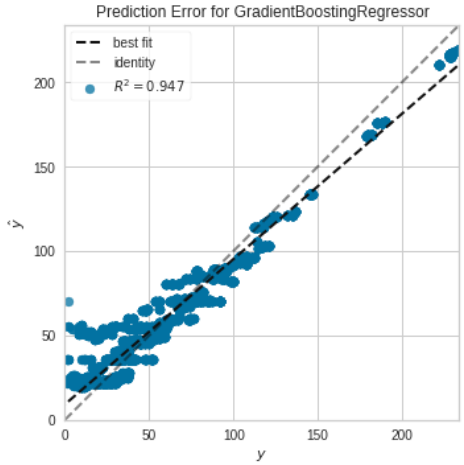
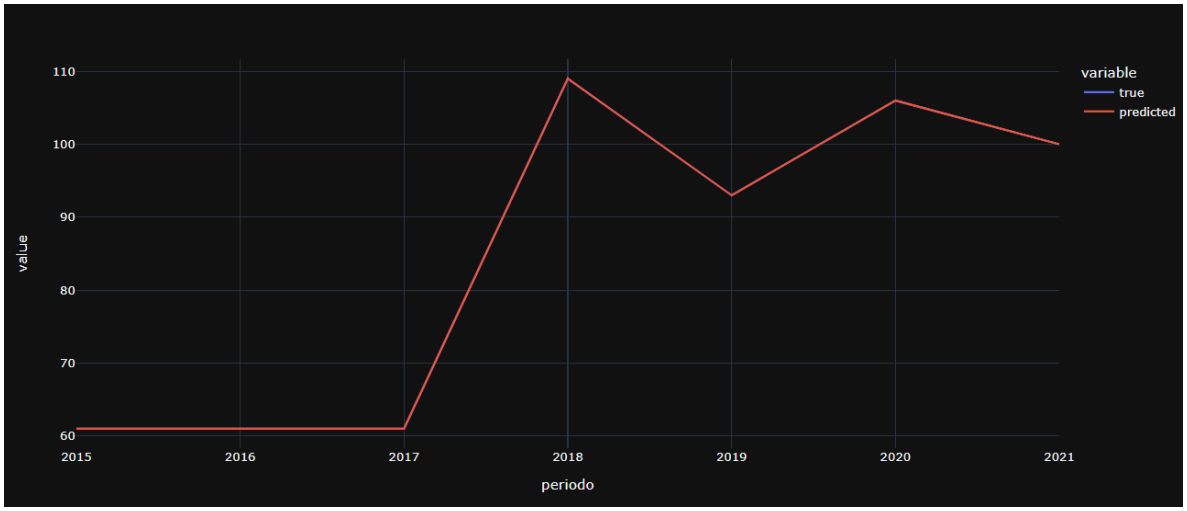


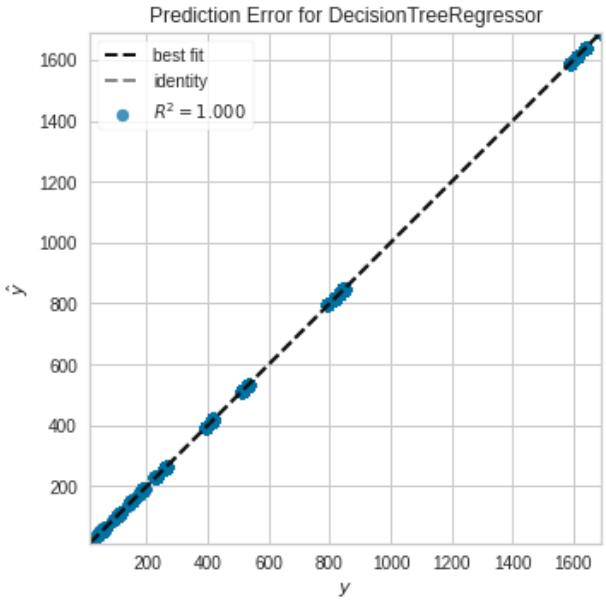
Figura 45. Cantidad de administrativos de 2018 a 2021

```
[ ]  dfBogota = predictionsAll[predictionsAll['NIVEL'] == 'Ejecutivo']
```

```
dfBogota2015 = dfBogota[dfBogota['YEAR'] == 2015]
t2015 = dfBogota2015['CantidadAdmin'].mean()
p2015 = dfBogota2015['Label'].mean()

dfBogota2016 = dfBogota[dfBogota['YEAR'] ==2016]
t2016 = dfBogota2016['CantidadAdmin'].mean()
p2016 = dfBogota2016['Label'].mean()

dfBogota2017 = dfBogota[dfBogota['YEAR'] == 2017]
t2017 = dfBogota2017['CantidadAdmin'].mean()
p2017 = dfBogota2017['Label'].mean()

dfBogota2018 = dfBogota[dfBogota['YEAR'] == 2018]
t2018 = dfBogota2018['CantidadAdmin'].mean()
p2018 = dfBogota2018['Label'].mean()

dfBogota2019 = dfBogota[dfBogota['YEAR'] == 2019]
t2019 = dfBogota2019['CantidadAdmin'].mean()
p2019 = dfBogota2019['Label'].mean()

dfBogota2020 = dfBogota[dfBogota['YEAR'] == 2020]
t2020 = dfBogota2020['CantidadAdmin'].mean()
p2020 = dfBogota2020['Label'].mean()

dfBogota2021 = dfBogota[dfBogota['YEAR'] == 2021]
t2021 = dfBogota2021['CantidadAdmin'].mean()
p2021 = dfBogota2021['Label'].mean()
```



	periodo	true	predicted
0	2015	71.0	71.0
1	2016	72.0	72.0
2	2017	74.0	74.0
3	2018	147.0	147.0
4	2019	145.0	145.0
5	2020	146.0	146.0
6	2021	144.0	146.0



De esta forma, se realizó la validación estática de los modelos y, como se pudo evidenciar, los modelos de *Random Forest Regressor* cuenta con un R2 de más del 90%.

Construcción del front y el backend

A continuación, se creó un backen en Flask. Este es un framework minimalista escrito en python que permite crear aplicaciones web rápidamente y con un mínimo número de líneas de código. La aplicación permite la carga de los modelos exportados desde Pycaret y, además, permite cargar los archivos de excel con la información necesaria de los datasets originales.

Figura 46. Aplicación de Flask

```
app = Flask(__name__)
CORS(app)
model_matriculados_posgrado = load_model('modelos/modeloMatriculadosPosgradoPycaret')
model_matriculados_pregrado = load_model('modelos/modeloMatriculadosPregradoPycaret')
model_graduados_posgrado = load_model('modelos/modeloGraduadosPosgradoPycaret')
model_graduados_pregrado = load_model('modelos/modeloGraduadosPregradoPycaret')
model_docentes = load_model('modelos/modeloDocentesPycaret')
model_admin = load_model('modelos/modeloAdministrativosPycaret')

PATH_MATRICULADOS_POSGRADO_CM = 'Periodo/matriculadosPosgradoCM.xlsx'
PATH_MATRICULADOS_PREGRADO_CM = 'Periodo/matriculadosPregradoCM.xlsx'
PATH_GRADUADOS_POSGRADO_CM = 'Periodo/graduadosPosgradoCM.xlsx'
PATH_GRADUADOS_PREGRADO_CM = 'Periodo/graduadosPregradoCM.xlsx'
PATH_DOCENTES_CM = 'Periodo/docentesCM.xlsx'
PATH_ADMIN_CM = 'Periodo/adminCM.xlsx'

matriculadosPosgradoCM = pd.read_excel(PATH_MATRICULADOS_POSGRADO_CM)
matriculadosPregradoCM = pd.read_excel(PATH_MATRICULADOS_PREGRADO_CM)
graduadosPosgradoCM = pd.read_excel(PATH_GRADUADOS_POSGRADO_CM)
graduadosPregradoCM = pd.read_excel(PATH_GRADUADOS_PREGRADO_CM)
docentesCM = pd.read_excel(PATH_DOCENTES_CM)
adminCM = pd.read_excel(PATH_ADMIN_CM)
```

Luego, se expusieron servicios de API REST para cada modelo. Uno de los requerimientos del cliente fue realizar la consulta solamente con el programa para matriculados y graduados, la unidad para los docentes y el nivel para los administrativos, es decir, no utilizar el conjunto de las variables, sino solamente aquellas mencionadas. Teniendo en cuenta que los modelos reciben todas las variables independientes para calcular la variable dependiente, se intentó cargar los dataset originales y hacer el filtrado directamente desde la aplicación, pero los archivos de excel son demasiado grandes para procesarlos en producción. En ese sentido, se sugiere la creación de una base de datos con estos archivos de forma que se pueda consolidar la información y hacer las consultas necesarias para las predicciones. Para solventar esta situación, se realizó un *group by* en los dataset originales con todas las combinaciones de las variables independientes excluyendo el programa, la unidad y el nivel, las cuales serían seleccionadas por el usuario desde el frontend. Esta es una solución parcial a este problema que excede los límites del presente trabajo debido a que se debería crear una base de datos y realizar el mapeo y los servicios necesarios del backend para traer la información necesaria para los modelos. La

versión actual permite consultar por programa, unidad y nivel para cada semestre, lo que refleja parcialmente el patrón encontrado por el modelo, pero no se corresponde estrictamente con los registros de los matriculados, graduados, docentes y administrativos para cada semestre o año. De igual forma, para consultar el histórico de matriculados, graduados docentes y administrativos, y compararlo con las predicciones, se creó un nuevo archivo de excel filtrado por periodo, nivel, programa y cantidad para matriculados y graduados, año, unidad y cantidad para docentes y año, nivel y cantidad para administrativos.

De esta manera, se expusieron los servicios para consultar los modelos seleccionados. Por ejemplo, para los matriculados en posgrado, se creó el servicio MatriculadosPosgradoPrograma(), el cual recibe un JSON por API REST con el nivel y el programa a consultar. Luego, se extrae del excel el histórico y se predice para cada semestre con todas las combinaciones de las variables independientes y se calcula la media. El backend responde al frontend con un JSON con las predicciones y el histórico del programa seleccionado.

Figura 47. Servicio de matriculados a posgrado por programa

```
@app.route('/api/matriculados_posgrado_programa', methods=['POST'])
@cross_origin()
def matriculadosPosgradoPrograma():
    data = request.get_json()
    print(data)
    df = pd.json_normalize(data)
    print(df)


posgradoGB = posgrado
print('posgrado GB')
print(posgradoGB)
nivel = df['NIVEL'].values[0]
print('nivel')
print(nivel)

posgradoNivel = posgradoGB[posgradoGB['NIVEL'] == nivel]
posgradoNivelCM = matriculadosPosgradoCM[matriculadosPosgradoCM['NIVEL'] == nivel]
print('posgrado nivel')
print(posgradoNivel)
print('posgrado nivel CM')
print(posgradoNivelCM)

programa = df['PROGRAMA'].values[0]
print('programa')
print(programa)
posgradoNivelPrograma = posgradoNivel[posgradoNivel['PROGRAMA'] == programa ]
posgradoNivelProgramaCM = posgradoNivelCM[posgradoNivelCM['PROGRAMA'] == programa ]
```

```
posgradoPrograma2009_1 = posgradoNivelPrograma
posgradoPrograma2009_1['PERIODO'] = '2009-1'
prediction = predict_model(model_matriculados_posgrado, data= posgradoPrograma2009_1)
mean2009_1 = prediction['Label'].mean()


posgradoPrograma2009_7 = posgradoNivelPrograma
posgradoPrograma2009_7['PERIODO'] = '2009-7'
prediction2 = predict_model(model_matriculados_posgrado, data= posgradoPrograma2009_7)
mean2009_7 = prediction2['Label'].mean()

posgradoPrograma2010_1 = posgradoNivelPrograma
posgradoPrograma2010_1['PERIODO'] = '2010-1'
prediction3 = predict_model(model_matriculados_posgrado, data= posgradoPrograma2010_1)
mean2010_1 = prediction3['Label'].mean()
```

```
response = pd.concat([
predictionDataframe, CM
])
response = response.to_json(orient="records")
parsed = json.loads(response)
response = json.dumps(parsed, indent=4)
return response
```

Siguiendo el mismo ejemplo, en el caso de las consultas con todas las variables independientes, se crea a su vez un servicio MatriculadosPosgrado() que recibe una petición con los parámetros correspondientes. Se extrae del excel el histórico y se predice para cada año como se puede ver en la siguiente figura. Este proceso se repitió para el resto del conjunto de datos.

Figura 48. Servicio de matriculados a posgrado

```
@app.route('/api/matriculados_posgrado', methods=['POST'])
@cross_origin()
def matriculadosPosgrado():
    data = request.get_json()
    print(data)

    nivel = data['NIVEL']
    print('nivel')
    print(nivel)
```

```
posgradoNivelCM = matriculadosPosgradoCM[matriculadosPosgradoCM['NIVEL'] == nivel]

print('posgrado nivel CM')
print(posgradoNivelCM)

programa = data['PROGRAMA']
print('programa')
print(programa)

posgradoNivelProgramaCM = posgradoNivelCM[posgradoNivelCM['PROGRAMA'] == programa ]



df20091 = pd.json_normalize(data)
df20091['PERIODO'] = '2009-1'
prediction20091 = predict_model(model_matriculados_posgrado, data=df20091)

df20097 = pd.json_normalize(data)
df20097['PERIODO'] = '2009-7'
prediction20097 = predict_model(model_matriculados_posgrado, data=df20097)

df20101 = pd.json_normalize(data)
df20101['PERIODO'] = '2010-1'
prediction20101 = predict_model(model_matriculados_posgrado, data=df20101)

df20107 = pd.json_normalize(data)
df20107['PERIODO'] = '2010-7'
prediction20107 = predict_model(model_matriculados_posgrado, data=df20107)



response = pd.concat([
predictionDataframe, CM
])
response = response.to_json(orient="records")
parsed = json.loads(response)
response = json.dumps(parsed, indent=4)
return response
```

Por su parte, en el frontend se construyó en Angularjs, el cual es un framework de código abierto para aplicaciones web construido en TypeScript y mantenido por Google. Este se utiliza para crear y mantener aplicaciones de una sola página basadas en el navegador y siguiendo la arquitectura Modelo Vista Controlador (MVC). Para cada conjunto de datos, se creó un formulario HTML con todas las variables independientes y los campos correspondientes para las consultas por programa, unidad o nivel. El ejemplo se puede ver en la siguiente figura.

Figura 49. Formulario HTML con todas las variables

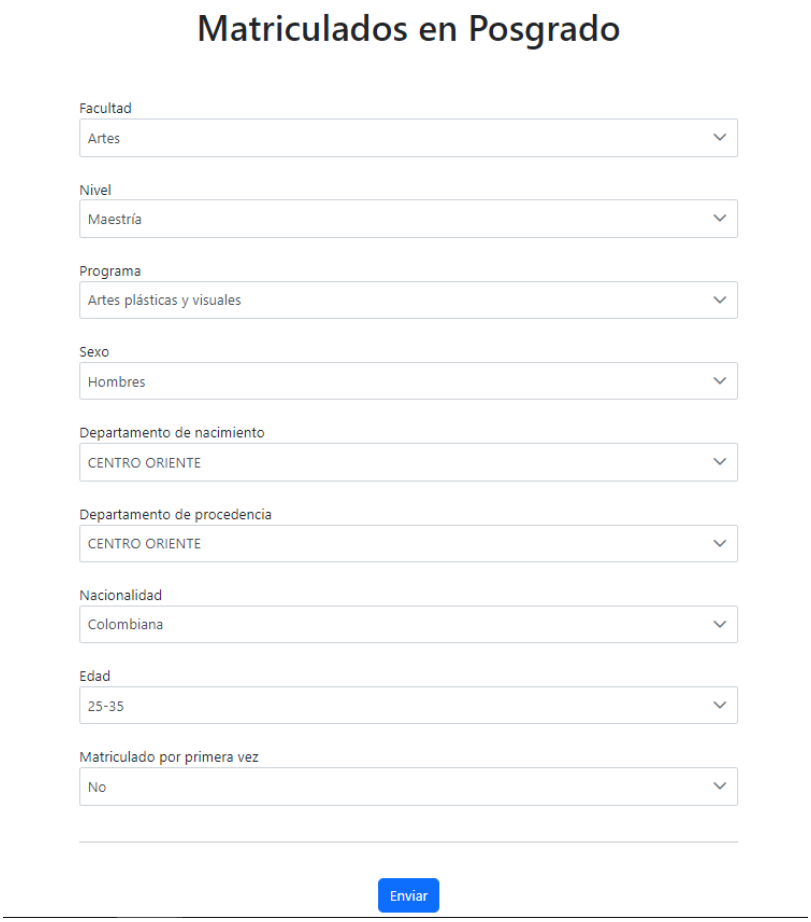


Figura 50. Formulario HTML para posgrado por programa

Los resultados se graficaron con Chart JS, una biblioteca JavaScript gratuita de código abierto para la visualización de datos, como se puede ver a continuación. Como ejemplo, se le preguntó al modelo la cantidad de matriculados en la Maestría de Arquitectura. La línea en azul corresponde al histórico mientras que la línea roja corresponde a las predicciones.

Figura 51. Visualización de los resultados

Los modelos se seleccionaron no solamente por sus métricas de predicción sino también por su facilidad de implementación. De esta manera, al utilizar la red neuronal de regresión para el caso de matriculados en pregrado, se obtuvieron inconsistencias en las predicciones debido a que al momento de normalizar o escalar los valores ingresados en el frontend, no se corresponden con la escala del entrenamiento por lo que se producen resultados inconsistentes. Por esta razón, se decidió utilizar todos los modelos elaborados con el algoritmo *Random Forest Regressor* de Pycaret, por su facilidad de integración con Flask. Además, esta librería también ofrece las gráficas de las métricas de evaluación. Otro requerimiento del cliente fue la necesidad de exportar los resultados en un archivo de excel. Esta funcionalidad también se desarrolló, de modo que los resultados de la consulta se extraen del JSON directamente al excel.

Figura 52. Métricas de los modelos

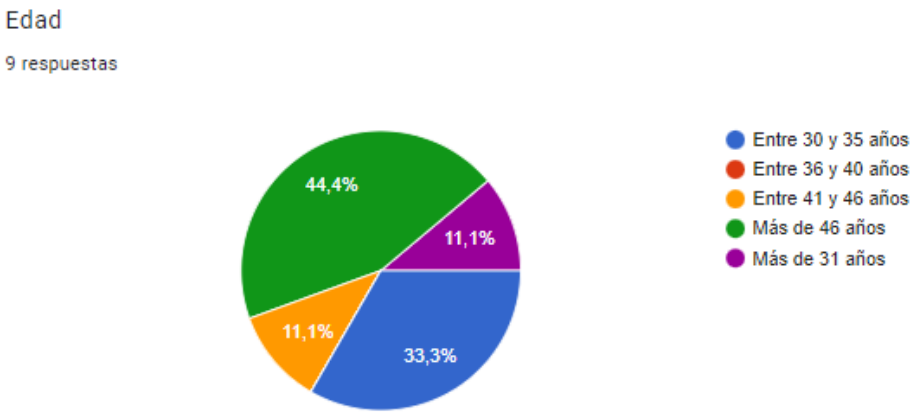
Validación dinámica

Siguiendo el ciclo de transferencia tecnológica de ingeniería, una vez construido el artefacto de software, se procede a realizar su validación en un contexto real con el fin de evaluar las variables que corresponden al impacto del mismo en la organización como solución a la problemática identificada. Así, siguiendo a R Bolaños [22], surge la necesidad de medir la percepción de utilidad de las predicciones a partir de la investigación cualitativa. De acuerdo con el autor, esta estudia el significado de las acciones humanas y la vida social como una interacción simbólica que se trata de manera explícita. [22, p. 30] De acuerdo con Meneses y Rodríguez [23], el investigador social no siempre puede acceder cuantitativamente a su objeto de análisis.. De esta forma, ellos plantean las preguntas subjetivas como aquellas en las que el ejercicio reflexivo de la persona reporta una información que no puede ser contrastada sino como un estado subjetivo autoinformado del que no existe ningún otro medio para acceder que el juicio del propio sujeto. [23, p. 12] Así, se determinó la encuesta como la metodología para evaluar la percepción de utilidad de los modelos predictivos y el cuestionario como el instrumento estandarizado para recoger información estructurada y cuantificable de la experiencia individual de las personas.

La población objetivo son los integrantes de la Oficina de Planeación y Estadística (OPE) de la UNAL. Se elaboró el consentimiento informado según los lineamientos éticos de ingeniería y el tratamiento de datos personales. Se recogió además información básica de los participantes como lo son la edad y el correo electrónico. Luego, se elaboraron 6 preguntas cerradas con respuestas ordenadas jerárquicamente, es decir, de tipo likert, para medir la importancia de conocer la tendencia del comportamiento poblacional de matriculados, docentes, graduados y administrativos en la planeación institucional, cómo se valoran los resultados de los modelos, cuál es el grado de precisión que se les atribuye y qué tan probable es que se utilice la herramienta en el trabajo diario. De igual manera, se elaboraron dos preguntas abiertas para conocer cómo el modelo ayudaría en la planeación estratégica y qué oportunidades de mejora se pueden realizar en el software.

Los resultados del cuestionario fueron los siguientes: en total, se contaron con 10 respuestas voluntarias de los integrantes de la OPE; en cuanto a las edad, se obtuvieron 9 respuestas, en las cuales 44% de las personas tiene más de 46 años, 11% tienen más de 31 años y entre 41 y 46 años y el 33% entre 30 y 35 años.

Figura 53. Edad



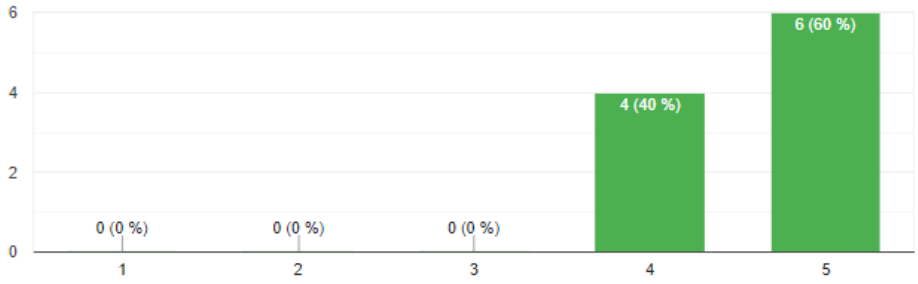
A la pregunta "¿Qué tan importante es conocer la tendencia del comportamiento poblacional en la Universidad Nacional de Colombia, sede Bogotá?", el 60% respondió con 5, es decir, muy importante, y el 40% respondió 4, es decir, importante. A la segunda pregunta, "¿Qué tan influyente es el comportamiento poblacional en la planeación presupuestaria institucional?", el 60% respondió 4 (importante) y el 40% 5 (muy importante). A la tercera pregunta, "¿Cómo valora los resultados generados por el modelo predictivo?", el 90% respondió 4 (buenos) y el 10% 5 (excelente). A la siguiente pregunta, "¿Qué grado de precisión considera que tienen los resultados de los modelos predictivos?", el 40 % respondió 3 (suficientemente buena) y el 60 % 4 (buena). A la última pregunta, "¿Qué tan probable es que haga uso del modelo predictivo en su trabajo?", el 10% respondió 2 (poco probable), el 70% respondió 4 (muy probable) y el 20% 5 (extremadamente probable).

Figura 54. Preguntas cerradas

1. ¿Qué tan importante es conocer la tendencia del comportamiento poblacional en la Universidad Nacional de Colombia, sede Bogotá?

Copiar

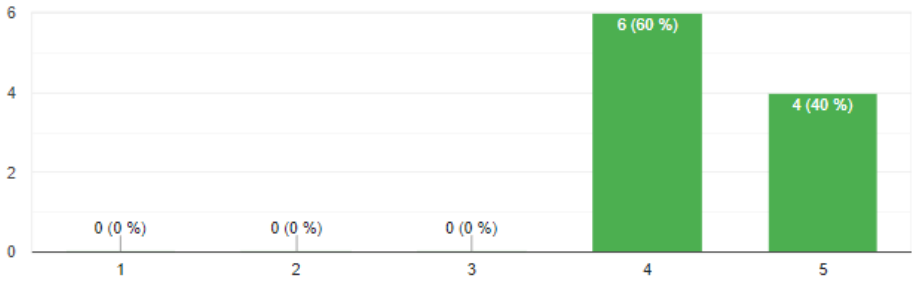
10 respuestas



2. En general, ¿Qué tan influyente es el comportamiento poblacional en la planeación presupuestaria institucional?

Copiar

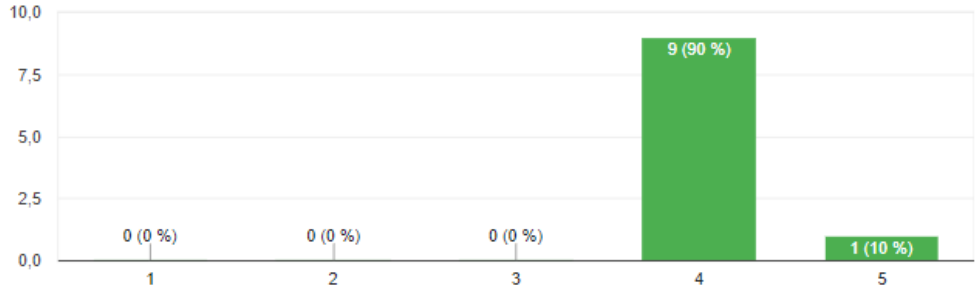
10 respuestas

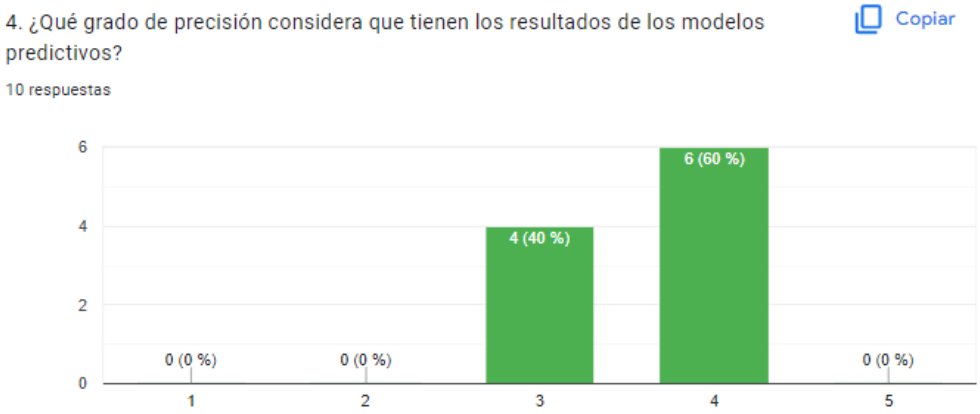


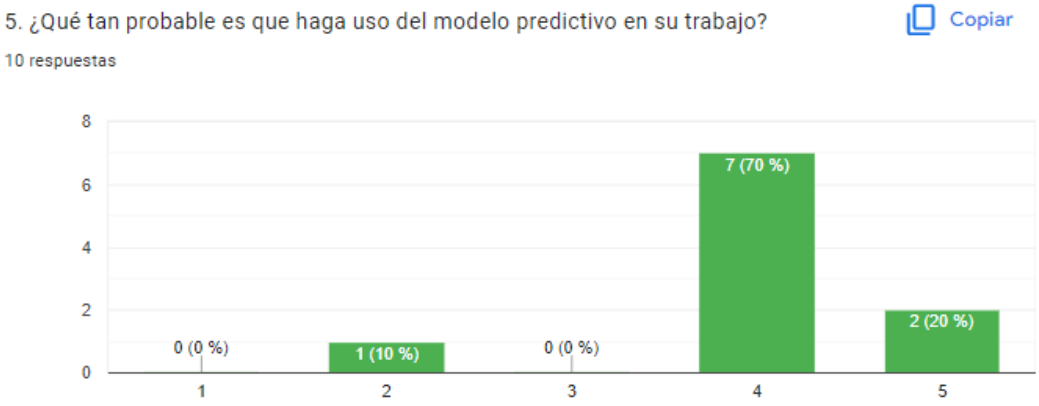
3. ¿Cómo valora los resultados generados por el modelo predictivo?

Copiar

10 respuestas







Por su parte, de las respuestas abiertas se pudieron establecer algunas categorías de acuerdo a su relación semántica. A la pregunta, "¿Cómo el modelo predictivo ayudaría en la planeación estratégica?", se determinó que la importancia de los modelos radica en la optimización de recursos físicos, humanos y financieros en la toma de decisiones. Así, los encuestados afirmaron lo siguiente:

> "El modelo es útil para realizar procesos de planeación financiera y de otro tipo de recursos necesarios para atender las necesidades de los diferentes actores y de la comunidad universitaria en general".
>
> "Claro que sí, sería una herramienta de gran ayuda para la toma de decisiones, en el que permitiría poder llegar a predecir y determinar los posibles impactos que se podrían generar en referencia a los objetivos misionales y requerimientos presupuestales en el desarrollo institucional a largo plazo, en su acreditación e infraestructura física y bienestar, la planta académica, docente y administrativa"

"Como herramienta para la toma de decisiones en el corto y mediano plazo, en temas como la optimización del recurso humano y financiero".
"Programación de recursos físicos, financieros, humanos".
"Ayudaría en la toma de decisiones y distribución del presupuesto".
"Para la proyección de nuevas inversiones".

Por otra parte, los encuestados resaltaron la importancia de prever la demanda de los servicios y plantear alternativas de acción en la planeación estratégica:

"Permitiría realizar proyecciones de demanda de servicios por parte de la comunidad universitaria en relación con la identificación de los clientes o usuarios potenciales y la demanda en temas de infraestructura, bienestar, calidad académica, entre otros aspectos".
"También permitiría ahorrar recursos de acuerdo a las predicciones realizadas o mostrar datos útiles para definir alternativas de acción con respecto a los procesos de la Universidad".

En cuanto a la última pregunta, "¿Qué oportunidades de mejora considera que puede tener la herramienta?", se pudieron identificar las siguientes categorías semánticas. Es importante identificar cuáles son las necesidades de uso de la herramienta desde la perspectiva del usuario, es decir, es indispensable facilitar la experiencia de consulta al usuario al momento de usar el artefacto:

"La herramienta debe darle información a los líderes de procesos, o jefes de dependencias, que sea de su valor para la toma de decisiones, así que deben indagar con estas partes interesadas qué información es importante para ellos, en la planeación y ejecución".
"Que el sistema pueda generar un listado de posibles preguntas partiendo de las variables que se determinan cuando se realicen dichas combinaciones de búsqueda y le permita al usuario tener la tranquilidad en la respuesta para la toma de decisiones".
"Considero que debe contener un instructivo para su consulta".

Por otra parte, es indispensable incluir variables exógenas a la universidad para ampliar el alcance de los modelos:

"Incluir variables exógenas y cualitativas que permitan dar respuesta a preguntas de tipo puntual por facultad o por proceso permitiendo a los interesados tener herramientas más

puntuales y decantadas para la toma de decisiones, también se podría realizar el ajuste de que no todos lo filtros de variables sean exigibles al buscar la proyección solicitada".
"Incluir otras variables exógenas que inciden en el número de matriculados, docentes y administrativos".
"Incluir variables exógenas y mediciones de contexto para comparar comportamientos y predicciones.".

Así mismo, es importante precisar el alcance y las restricciones del proyecto para evitar confusiones:

"Es importante aclarar que el modelo que están presentando para la predicción en el crecimiento poblacional es en la Sede Bogotá, como estudio. No están incluidas todas las variables en las categorías que definen la matriz Docente, el cual no están incluidas los docentes ocasionales, (generando un sesgo en la información). Indicar que corresponde únicamente a los docentes de planta".
"Inclusión de criterios restrictivos para el crecimiento poblacional (por ejemplo: la restricción de ampliación de la planta docente y administrativa)".

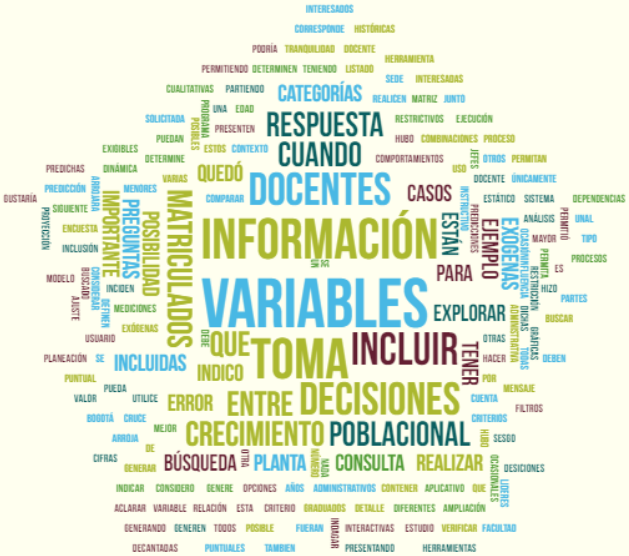
Por último, se propone además ampliar el alcance de los modelos con el cruce de variables entre ellos:

"Explorar la posibilidad de realizar análisis entre las diferentes categorías (ejemplo: relación entre matriculados y graduados o entre matriculados y docentes)".
"Considerar el detalle en el cruce de información para el uso de esta en la toma de decisiones. "
"Explorar la posibilidad de que se determine la variable de mayor influencia en el crecimiento poblacional."

Estas categorías, se puede apreciar igualmente en las siguientes nubes de palabras de las dos preguntas abiertas.

Figura 55. Nube de palabras, pregunta 6

Figura 56. Nube de palabras, pregunta 7



ANÁLISIS DE RESULTADOS

La Oficina de Planeación y Estadística (OPE) de la Universidad Nacional de Colombia (UNAL) es la encargada de coordinar e integrar los procesos de planeación de la sede de Bogotá. Así mismo, asesora las diferentes facultades y dependencias en la implementación del Plan Estratégico Institucional, el Plan Global de Desarrollo y el Plan de Acción Institucional. Desde el enfoque del modelo Biopsicosocial y cultural de la Universidad El Bosque, el *medio* en el que se inscribe la solución de ingeniería es la Universidad Nacional dentro de la *cultura* institucional definida por su misión y visión propias. Los *hábitos* de los integrantes de la OPE se corresponden a la recopilación de datos y al análisis descriptivo y de diagnóstico de los mismo con *artefactos* muy importantes para la planeación estratégica como lo son los tableros interactivos y la Plataforma de Registro de los Informes de Gestión (PRIG). Una de las *creencias* base de la cultura organizacional es crear un sistema de información unificado para toda la sede Bogotá de modo que se pueda cruzar toda la información para mejorar los procesos de toma de decisiones.

Actualmente, se han realizado predicciones lineales en excel del número de estudiantes, sin embargo, su índice de correlación no es muy preciso y no se están tomando en cuenta las características poblaciones de los estudiantes. Además, no se han ampliado estas técnicas al estudio de otros conjuntos de datos como el crecimiento poblacional de matriculados, graduados, docentes y administrativos. La presente solución de ingeniería se propuso dar el primer paso hacia el análisis predictivo de un conjunto de datos disponibles. El *medio* y los *actores* del modelo biopsicosocial son los mismos, sin embargo, se propone transformar los hábitos de los integrantes de la OPE para utilizar herramientas de analítica predictiva con la finalidad de apoyar la toma decisiones en la asignación de recursos disponibles e identificar riesgos y oportunidades. Esto afecta las dimensiones económicas, sociales y culturales del entorno universitario ya que facilita la planeación estratégica de la sede.

Así, se cumplió el objetivo general de desarrollar modelos predictivos de las tendencias del crecimiento poblacional de los matriculados, docentes, graduados y administrativos mediante algoritmos de aprendizaje automático. Siguiendo la metodología ASUM-DM, se realizó en primera instancia un análisis técnico que se consolidó en el acta de constitución del proyecto en el que se establecieron los costos, los requerimientos de alto nivel (principalmente que los modelos tuvieran

más del 80% de precisión en sus métricas y que se pudieran consultar los resultados en una interfaz web) y los riesgos: los modelos no generan el valor esperado para el negocio o no se cumple con el cronograma establecido. Se realizó, además, de manera iterativa y de la mano del cliente y el director del proyecto de grado, el proceso de entendimiento del negocio, la preparación de los datos y la construcción de los modelos. Esto se puede evidenciar en los reportes de preparación y construcción de los modelos entregados al cliente, las cuales incluyen implícitamente la validación en la academia (anexo 8). De igual forma, se realizó la evaluación de los modelos con datos no incluidos en el entrenamiento cuyos resultados se pueden evidenciar en la validación estática del ciclo de transferencia tecnológica (CTT). Una vez finalizados los modelos, se construyó un frontend y un backend para realizar consultas y se desplegó la solución en un ambiente controlado. Por último, de acuerdo a la metodología, se llevó a cabo un primer acercamiento a la transferencia de conocimiento operacional con los usuarios. Esto permitió, en principio, efectuar una validación dinámica del artefacto en el marco del CTT para evaluar la percepción de utilidad del software por parte de los integrantes de la OPE. A partir de la investigación cualitativa y la metodología de encuesta, el cuestionario permitió cuantificar el sentido subjetivo del autoinforme de los participantes. De esta forma, se tuvieron resultados positivos en los que el 60% de la muestra considera importante y, el 40% muy importante, conocer el crecimiento poblacional en la planeación estratégica, el 90% considera buenos los resultados generados por el artefacto y, el 10%, excelentes, el 60% considera que la precisión de los modelos es buena y el 40% restante suficientemente buena y el 90% de la población considera que es muy probable que utilice los modelos en su trabajo en la OPE. En cuanto a las preguntas abiertas, se evidenció que la importancia de los modelos radica en la optimización de recursos físicos, humanos y financieros y poder prever la demanda de los servicios ofrecidos por la universidad, así como plantear alternativas de acción en caso de riesgos.

En cuanto a las limitaciones del proyecto, en la construcción del backend, no se logró realizar las predicciones con los datos reales por semestre o año sino con todas las combinaciones de las variables independientes para mostrar el patrón identificado por el algoritmo. Debido al tamaño de los archivos de excel, no se pudieron cargar en producción y extraer los datos reales necesarios. Esto se debió a que la construcción de una base de datos y las consultas necesarias excede el tiempo previsto por el cronograma teniendo en cuenta la curva de aprendizaje del desarrollo de aplicaciones en el framework Flask. Por otra parte, para sorpresa de todos los involucrados, los modelos de administrativos y docentes no generan el valor esperado al predecir una línea constante. Esto no se

debe a un error en la creación o evaluación de los mismos, sino a que el histórico de los datos es constante y el enfoque de árboles de decisión para todos los niveles y unidades no es capaz de modelar estas pequeñas diferencias. De esta forma, se propone utilizar series de tiempo para cada unidad o nivel específico como una alternativa al presente enfoque.

Según lo descubierto en el estado del arte, las universidades hoy en día operan en un entorno competitivo y complejo. El análisis de la información disponible ayuda en la creación de estrategias para evaluar el rendimiento de estudiantes y profesores, aumentar la tasa de retención, mejorar el marketing y la eficacia general de la organización [6, p.6] Como se mencionó desde el modelo biopsicosocial y cultural, el hábito de usar herramientas de predicción en la planeación institucional afecta las dimensiones sociales, económicas y culturales de la universidad y, en última instancia, puede mejorar la calidad de vida de las personas vinculadas a ella. Para este trabajo se encontró que los árboles de decisión aleatorios (*Random Forest Regressor*) son óptimos para aprender de un conjunto de datos histórico como el enunciado en este trabajo. El clasificador equivale a una serie de declaraciones IF-THEN que modelan la estructura de las variables independientes para predecir la variable dependiente. Los modelos generados en su mayoría alcanzaron una R2 del 90%.

Otro descubrimiento importante del presente proyecto fueron las herramientas de *automated machine learning*. H. Zeineddine, U. Braendle and A. Farah [15] recomiendan a los investigadores utilizar estas técnicas pues incrementan la productividad de los científicos de datos. AutoML escoge el mejor modelo de clasificación o regresión en un grupo de algoritmos y, definitivamente, facilitó la búsqueda de los algoritmos en este estudio. Además, la librería de Pycaret ofrece herramientas útiles como la creación de las gráficas de las métricas y facilidad en la exportación e implementación de los modelos.

La analítica de datos tiene una amplia gama de metodologías que se clasifican, en términos generales, en descriptivos, diagnósticos, predictivos y prescriptivos. La visualización de datos futuros a partir de predicciones es una herramienta esencial en la planeación de cualquier organización ya que permite reducir la incertidumbre en un entorno siempre cambiante, generar valor comercial, prever problemas y oportunidades, optimizar la asignación de recursos y la gestión y, finalmente, mejorar la calidad de vida de las personas. En este sentido, el presente proyecto cumplió el objetivo de generar modelos predictivos del crecimiento poblacional de los matriculados, docentes, administrativos y

graduados con el fin de dar un primer paso para inculcar en los integrantes de la OPE el hábito de utilizar analítica predictiva y prescriptiva en la planeación.

CONCLUSIONES

De acuerdo con los lineamientos del programa de Ingeniería de Sistemas de la Universidad El Bosque, el ciclo de transferencia tecnológica permite efectuar una transformación de una problemática. El modelo biopsicosocial y cultural, por su parte, permite identificar claramente el problema y proponer una solución a pastor de un artefacto de software. El presente trabajo logró realizar los modelos predictivos de las tendencias del crecimiento poblacional de matriculados, docentes, graduados y administrativos mediante algoritmos de *machine learning* para apoyar la planeación estratégica institucional que se lleva a cabo en la Oficina de Planeación y Estadística de la sede Bogotá de la Universidad Nacional de Colombia. Siguiendo la metodología ASUM-DM se cumplieron los objetivos específicos de determinar los requerimientos de información y las fuentes de datos, definir y aplicar los algoritmos y validar la percepción de utilidad de los modelos en el trabajo de los integrantes de la OPE. Este trabajo es útil y novedoso en el sentido que da el primer paso para cambiar el hábito de los tomadores de decisiones en el uso de la analítica descriptiva y diagnóstica hacia la predictiva y prescriptiva con el fin de generar valor en la gestión institucional global a partir de la reducción de la incertidumbre y los riesgos y la capitalización de las oportunidades.

Por otra parte, se encontró que las técnicas de AutoML aceleran bastante los procesos de selección de los algoritmos y su implementación en ambientes de producción. Además, para todos los modelos se seleccionó el *Random Forest Regressor* por su capacidad de construir la estructura adecuada para predecir los valores de la variable dependiente a partir de las variables independientes con más de un 90% de precisión.

Como líneas de investigaciones futuras, se propone mejorar continuamente los modelos con los nuevos datos que llegan a la OPE, la inclusión de variables exógenas en los modelos provenientes, por ejemplo, de las bases de datos del Ministerio de Educación, mejorar la experiencia de usuario del artefacto de modo que sea fácil realizar las consultas que realmente necesita, logrando así exitosamente una transferencia adecuada del conocimiento, y, por último, permitir el cruce de los

diferentes conjuntos de datos matriculados-graduados, graduados-docentes, etc.

6. LECCIONES APRENDIDAS

Teniendo en cuenta que no se había trabajo en un proyecto relacionado con análisis de datos no se tenía conocimiento exhaustivo ni de las técnicas ni de las metodologías existentes para el desarrollo de proyectos de este tipo, con lo cual, en principio, se pensó en trabajar con la metodología SCRUM, la cual es una metodología que contempla el desarrollo ágil, pero no es la más apropiada para el desarrollo de software en proyectos basados en datos. Además, como no se cuenta con amplios conocimientos ni experiencia en el campo de la ciencia de datos, buscamos la asesoría de un docente especializado, el cual sugirió también trabajar con técnicas de auto machine learning. Tras el proceso de depuración, se inicia la fase de análisis exploratorio, con lo cual no resulta fácil visualizar la correlación de más de dos variables debido a la cantidad de datos que se involucran, el cliente nos manifiesta que la visualización presentada no es muy diciente y que resulta difícil inferir información útil al respecto, con lo cual, tras realizar varias iteraciones sobre el producto, se realizar nuevamente una nueva gráfica que permite mostrar de una manera más eficiente los datos que se quieren interpretar. Por otra parte, al momento de construir los modelos, se descubrió que las series de tiempo no eran viables para este proyecto dado que se debía realizar un modelo para cada programa lo cual era insostenible en el tiempo debido su implementación en el backend. Además, cuando se utilizó la herramienta Autokeras para definir arquitecturas de redes neuronales que más se ajusten al conjunto de datos, los resultados no fueron satisfactorios ni lo suficientemente buenos comparados con las demás herramientas. Igualmente, una de las dificultades más grandes fue el manejo del cronograma al momento de estimar los tiempos de trabajo, por lo que se tuvo que actualizar constantemente para la creación y evaluación de los modelos y el desarrollo de los reportes del cliente. Finalmente, se adquirió un conocimiento inicial de la metodologías ágil ASUM-DM y de las herramientas herramientas de análisis de datos como Pycaret, Autokeras, SciKit learn, Tensorflow y Flask. Esto fue muy provechoso para la vida profesional ya que nos permitió ampliar nuestro conocimiento del desarrollo de software tradicional hacia los sistemas inteligentes.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] C. N. Miranda, T. J. Escobar y U. C. Escobar. Universidad el Bosque, una historia en construcción. Bogotá: Ediciones El Bosque, 2009.

[2] J. A. Montaña. "El modelo biopsicosocial y cultural para ingenierías: de la relación médico paciente a la relación sociológica del ingeniero con la comunidad". Departamento de Humanidades, Universidad el Bosque.

[3] Oficina de Planeación y Estadística. (2021, oct, 21). Quienes somos. [Online]: http://planeacion.bogota.unal.edu.co/quienes_somos/

[4] Oficina de Planeación y Estadística. (2021, 25,10). PRIG. [Online]: http://planeacion.bogota.unal.edu.co/prig/

[5] IBM. Analytics Solutions Unified Method (ASUM). Delivery Process: ASUM-DM. [Online]: http://i2t.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asumDM_Teaser/deliveryprocesses/ASUM-DM_8A5C87D5.html/

[6] B. Albreiki, N. Zaki and H. Alashwal, H. "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques." Education Sciences, 11, 9, 2021.

[7] H. Pallathadka, A. Wenda, E. Ramirez-As´ıs, M. As´ıs-L´opez, J. Flores-Albornoz and K. Phasinam. "Classification and prediction of student performance data using various machine learning algorithms." Materials Today: Proceedings.

[8] H. K.D. Menon and V. Janardhan, "Machine learning approaches in education." Materials Today: Proceedings, 43, 3470–3480, 2021.

[9] Universidad de Valladolid, B. "Ciclo de vida de los datos". [Online]: https://biblioguias.uva.es/datos-investigacion/plan-gestion-datos

[10] K. T Chui, D. C. L. Fung, M. D. Lytras and T. M. Lam. "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm." Computers in Human Behavior, 107, 105584, 2020.

[11] E. Gothai, R. Thamilselvan, R. Rajalaxmi, R. Sadana, A. Ragavi and R. Sakthivel. "Prediction of covid-19 growth and trend using machine learning approach." Materials Today: Proceedings.

[12] S. Tuli, S. Tuli, R. Tuli, and S. Gill. "Predicting the growth and trend of COVID- 19 pandemic using machine learning and cloud computing." Internet of Things, 11, 100222, 2020.

[13] X. Pang, C. B. Forrest, F. Lˆe-Scherban and A. J. Masino. "Prediction of early childhood

obesity with machine learning and electronic health record data." International Journal of Medical Informatics, 150 104454 (2021).

[14] K. van Mens, et al. "Predicting future suicidal behaviour in young adults, with different machine learning techniques: A population-based longitudinal study." Journal of affective disorders, 271, 169-177 (2020).

[15] H. Zeineddine, U. Braendle and A. Farah. "Enhancing prediction of student success: Automated machine learning approach." Computers & Electrical Engineering, 89, 106903 (2021).

[16] P. Dabhade, R. Agarwal, K. Alameen, A. Fathima, R. Sridharan and G. Gopakumar. "Educational data mining for predicting students' academic performance using machine learning algorithms." Materials Today: Proceedings.

[17] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt and E. Cascallar. "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation." Computers and Education: Artificial Intelligence, 2, 100018 (2021).

[18] J. Baijens, T. Huygh and R. Hemls. "Establishing and Theorising Data Analytics Governance: a Descriptive Framework and a VSM Based View", Journal of Business Analytics, DOI: 10.1080/2573234X.2021.1955021

[19] P. S. Deshpande, S. C. Sharma and S. K. Peddoju. "Predictive and Prescriptive Analytics in Big-data Era". In: Security and Data Storage Aspect in Cloud Computing. Studies in Big Data, vol 52. Springer, Singapore. https://doi-org.ezproxy.unbosque.edu.co/10.1007/978-981-13- 6089-3_5

[20] P. V. Britos. "Evaluación comparativa de las metodologías Team Data Science Process TDSP y Analytics Solutions Unified Method for Data Mining ASUM-DM desde la perspectiva de la ciencia de datos Giuliana Fois Gustavo Andrés Agüero Crovella". Investigación Formativa en Ingeniería (2020): 264.

[21] IBM. ¿Qué es machine learning? [Online] Available: https://www.ibm.com/co-es/analytics/machine-learning

[22] Auto-sklearn. Manual. [Online] Available: https://automl.github.io/auto-sklearn/master/#manual

[23] R. Bolaños. "La investigación cualitativa en las ciencias de la administración: aproximaciones tóricas y metodológicas". Revista nacional de administración, 8(1), 25-45, enero-junio, 2017.

[24] J. Meneses and D. Rodríguez. El cuestionario y la entrevista. Barcelona: Universitat Oberta de Catalunya, 2011.. https://femrecerca.cat/meneses/publication/cuestionario-entrevista

11. ANEXOS

Anexo 1. Modelo Biopsicosocial del problema

Anexo 2. Modelo Biopsicosocial de la solución

Anexo 3. Estructura de desglose de trabajo

Anexo 4. Cronograma

Anexo 5. Análisis técnico: plan de costos y riesgos

Anexo 6. Acta de constitución del proyecto

Anexo 7. Data set de matriculados en posgrado

Anexo 8. Reportes de preparación y construcción de modelos