Universidad de Valladolid

Escuela de Ingeniería Informática de Valladolid

Máster en Ingeniería Informática – Especialidad Big Data

Trabajo Fin de Máster

Análisis de datos y aprendizaje automático del proceso de admisión de la Universidad Nacional Autónoma Honduras para la región sur del país del 2006 al 2019

Autor:

Alex Dario Flores Aplicano

Tutor:

Quiliano Isaac Moro

“No puedo enseñar nada a nadie. Solo puedo hacerles pensar”

- Sócrates

Agradecimientos

En primer lugar, un agradecimiento enorme al programa de Beca Iberoamérica + Asia, el cual me ha permitido ser partícipe de esta experiencia rica en conocimientos teóricos y prácticos a través del estudio de este máster, así como conocer nuevas personas y sus diversas culturas.

A mi familia, quienes siempre han apoyado cada decisión que he tomado sin importar la distancia.

A mis amigos, quienes, a pesar de no siempre estar con ellos en todas sus locuras, siempre piensan en mí para acompañarlos.

Resumen

Toda universidad pública tiene responsabilidades con la sociedad, por lo que debe apuntar su labor a ofrecer el mejor servicio posible al ciudadano[1] proporcionando el recurso humano de más alto nivel académico[2]. Esto se asegura con un proceso de admisión transparente, eficiente, igualitario, de calidad y con equidad[3]. Por ello se analizan los datos recopilados en la etapa de inscripción y realización de la prueba de admisión de la Universidad Nacional Autónoma de Honduras (UNAH) en la región sur del país del año 2006 al 2019, utilizando las metodologías de minería de datos SEMMA y CRISP-DM para detectar con anticipación a realizar dicha prueba, los aspirantes vulnerables a **no** ser admitidos en la universidad.

Se comienza por una adaptación de los datos, luego se separan los datos en una parte para la creación del modelo y otra para la validación, para así obtener resultados libres de sesgos siguiendo un esquema de validación cruzada. Se prueban varios algoritmos de clasificación del modelo pertenecientes a las librerías sklearn y xgboost

El mejor modelo obtiene una precisión del 66% con una especificidad del 61% para determinar los aspirantes que no serán **admitidos** a la UNAH. Además, se adiciona un cuadro de mando para dar seguimiento a los procesos de admisión.

Abstract

Every public university has responsibilities to society, so it must aim its work to offer the best possible service to the citizen [1] by providing the highest academic level human resource [2]. This is ensured with a transparent, efficient, egalitarian, quality, and fair admission process [3]. For this reason, the data collected in the stage of enrollment and completion of the admission test of the National Autonomous University of Honduras (UNAH) in the southern region of the country from 2006 to 2019 is analyzed, using the data mining methodologies SEMMA and CRISP-DM to detect in advance to carry out said test, applicants vulnerable to not being admitted to the university.

It begins with an adaptation of the data, then the data is separated into one part for the creation of the model and another for the validation, to obtain bias-free results following a cross-validation scheme. Various model classification algorithms belonging to the sklearn and xgboost libraries are tested

The best model obtains an accuracy of 66% with a specificity of 61% to determine applicants who will not be admitted to UNAH. In addition, a dashboard is added to monitor the admission processes.

Índice de Contenidos

Índice de figuras

Índice de tablas

índice de ecuaciones

Capítulo 1

1. Introducción

1.1. Introducción

En los últimos años, las empresas han estado recopilando datos de sus procesos con el fin de analizarlos, buscando tendencias y comportamientos que les permita volverse más competitivas en el mercado. Las universidades, como instituciones destinadas a la enseñanza superior, buscan en su visión otorgar títulos y conocimientos de calidad a sus estudiantes en las diferentes ramas del saber. Por ello están constantemente recolectando información de sus procesos administrativos y de enseñanza, tal que con los datos se puedan realizar análisis que les permitan generar estrategias de mejora continua sobre su actividad ante la población. Uno de los procesos más delicados de toda universidad consiste en la admisión de los aspirantes con perfil de egresado de educación media que intentan ingresar a ella.

El proceso de admisión, más que una manera de seleccionar los aspirantes que serán admitidos para ingresar a las universidades, consiste en la generación de un diagnóstico sobre el nivel de conocimientos académicos con que cuentan los aspirantes, previos a ingresar a realizar estudios universitarios [4]. Esto orientado a promover transformaciones dentro de las estrategias que permitan mejorar el nivel de calidad educativo en la educación superior [3]. Por tanto, el reto que han de afrontar las universidades será el de relacionar los datos que han estado recopilando mediante la inscripción y realización de las pruebas de admisión, para detectar de forma temprana los aspirantes que corren el riesgo de no ser admitidos por no contar con los requisitos mínimos de conocimientos previos, así como los factores que involucran este posible resultado.

Se plantea el uso de los conocimientos adquiridos en las asignaturas del Máster en Ingeniería Informática de la Universidad de Valladolid, para el diseño de modelos de aprendizaje automático y de visualización sobre los datos del proceso de admisión de la Universidad Nacional Autónoma de Honduras (UNAH) para la región sur del país del 2006 al 2019, como apoyo a las estrategias que genera la Dirección del Sistema de Admisión (DSA), mediante el uso de metodologías que permitan comprender el negocio y los datos.

1.2. Motivación

El ingreso a la universidad en la actualidad es un objetivo importante para la población de estudiantes que egresan de los centros de educación media, dada la gran cantidad de beneficios laborales y de conocimientos que genera. Sin embargo, es un punto crítico debido a la gran cantidad de información que los aspirantes deben considerar (p. ej. oferta académica, criterios de selección, niveles de conocimientos previos) para ingresar.

Las universidades suelen contar con servicios profesionales de apoyo estudiantil para orientar a los aspirantes durante el proceso de admisión y aumentar sus probabilidades de éxito. También existen herramientas que aplican aprendizaje automático (p. ej. myKlovr [5][6]) para proporcionar asesoramiento que permita a los aspirantes la admisión a universidades, basado en patrones comunes de su perfil en comparación al de otros estudiantes que lograron ingresar con éxito a un centro de educación superior.

Actualmente, no existe solución en el mercado que permita a las universidades detectar con anticipación el grupo de aspirantes vulnerables a reprobar la prueba de admisión para ingresar a estudios universitarios. Esto implica, muchas oportunidades para la aplicación de técnicas de aprendizaje automático sobre los datos que se recopilan.

Identificar de manera temprana los aspirantes vulnerables a reprobar la prueba de admisión, permitirá a los responsables del sistema de admisión de las universidades:

- Mejorar sus estrategias para la captación de aspirantes.
- Proponer programas de preparación para la prueba de admisión a este grupo específico de aspirantes.
- Alertar a los centros de educación media las dificultades que presentan los aspirantes que reprueban el examen de admisión.

1.3. Objetivo General

Analizar los datos del proceso de admisión a la Universidad Nacional Autónoma de Honduras (UNAH) en la región sur del país, recopilados mediante la aplicación de la Prueba de Aptitud Académica (PAA) por la Dirección del Sistema de Admisión (DSA), para la detección temprana de aspirantes vulnerables a no ser admitidos en la universidad.

1.3.1. Objetivos específicos

1) Implementar las metodologías para la gestión de proyectos de Minería de Datos (Data Mining); SEMMA (Sample, Explore, Modify, Model, Assess) y CRISP-DM (Cross-Industry Standard Process for Data Mining) para comprender la necesidad del negocio y los datos a analizar.
2) Proponer un modelo predictivo de apoyo a las estrategias generadas para el mejoramiento continuo del proceso de admisión.
3) Crear un cuadro de mando (*dashboard*) que permita a la DSA realizar seguimiento a los procesos de admisión y apoyo a la toma de decisiones.

1.4. Estructura de capítulos

El documento está estructurado en capítulos de la siguiente manera:

- **Capítulo 2**: se describe el marco teórico y estado del arte, los antecedentes e historia de la UNAH, organización jerárquica y el proceso de admisión que el aspirante debe superar para ser **admitido** en la UNAH. Además, conceptos importantes sobre análisis de datos y aprendizaje automático.
- **Capítulo 3**: describe la metodología implementada, la cual es una combinación entre CRISP-DM y SEMMA ambas partes del paradigma KDD.
- **Capítulo 4**: en esta parte se desarrolla la metodología planteada, en ella se describe el proceso desde la comprensión de los datos hasta el despliegue del modelado final.
- **Capítulo 5**: diseño del cuadro de mando (*dashboard*) para el seguimiento de los procesos de admisión y apoyo a la toma de decisiones.
- **Capítulo 6**: presentación de las conclusiones y trabajo futuro.

Este trabajo contiene, al final del todo, anexos con tablas y figuras que se consideran importantes durante el procesamiento de los datos y generación de los modelos.

Capítulo 2

2. Marco Teórico y Estado del arte

2.1. Antecedentes

La Universidad Nacional Autónoma de Honduras (UNAH) es una institución autónoma, laica y estatal de la república. Universidad con carácter jurídico, encargada de gestionar la educación superior del país en los grados de licenciatura, maestría, segunda especialidad y doctoral.

2.1.1. Historia de la UNAH

En diciembre del año 1845, la primera institución de estudios superiores con carácter privado que recibió reconocimiento del gobierno del país fue "La Sociedad del Genio Emprendedor y del Buen Gusto", bajo iniciativa de José Trinidad Reyes quien ejerció como Rector, en compañía de Máximo Soto, Alejandro Flores, Miguel Antonio Rovelo, Yanuario Girón y Pedro Chirinos. En marzo del año 1846, el gobierno del país reconoce a la sociedad como "Academia Literaria de Tegucigalpa", la cual, posteriormente en septiembre del año 1847 bajo convenio entre José Trinidad Reyes y Juan Lindo (presidente de la república, 1847 - 1852), cambia al título de Universidad Central del Estado, ubicada en Tegucigalpa. En octubre del año 1957, la Junta Militar del Gobierno constituida por Héctor Caraccioli y Roberto Gálvez Barnes, consideran a la Universidad Central del Estado como Universidad Nacional Autónoma de Honduras. Esta autonomía, brinda a la universidad libertad de cátedra, estudio, investigación y vinculación con la sociedad, así como potestad para elegir a sus autoridades, emitir normas reglamentarias, generar políticas de ingreso, permanencia

y egreso de los estudiantes sin influencia política del estado. Al año 2020 han transcurrido 173 años desde que se inauguró dicha universidad [7].

En la actualidad, la UNAH cuenta con Centros Regionales [8], Centros de Aprendizaje de Educación a Distancia (CRAED) [9] y Telecentros Universitarios [10] para modalidad virtual ubicados estratégicamente en Honduras, como se muestra en la Tabla 1:

Tabla 1 - Centros de Estudios de la UNAH en Honduras

Región	Departamentos de Honduras	Centros de Estudios
Región Occidental	Ocotepeque Copán Lempira	Centro Universitario Regional de Occidente (UNAH – CUROC) CRAED – La Entrada Copán Telecentro – Ocotepeque Telecentro – Gracias
Región Noroccidental	Cortés Santa Barbara Yoro	UNAH Valle de Sula (UNAH – VS) Centro Tecnológico del Valle de Aguan (UNAH – TEC AGUÁN) CRAED – El Progreso Telecentro – Choloma Telecentro – Puerto Cortés
Región Nororiental	Atlántida Colón Gracias a Dios Islas de la Bahía	Centro Universitario Regional de Litoral Atlántico (UNAH – CURLA) CRAED – Tocoa Telecentro – Roatán
Región Centro-Occidental	Intibucá Comayagua La Paz	Centro Universitario Regional del Centro (UNAH – CURC) CRAED – Siguatepeque
Región Centro-Oriental	Francisco Morazán El Paraíso Olancho	Ciudad Universitaria (UNAH – CU) Centro Tecnológico de Danlí (UNAH – TEC Danlí) Centro Universitario Regional Nororiental (UNAH – CURNO) CRAED – El Paraíso

Región	Departamentos de Honduras	Centros de Estudios
		CRAED – Tegucigalpa CRAED – Juticalpa
Región Sur	Choluteca Valle	Centro Universitario Regional del Litoral Pacífico (UNAH – CURLP) CRAED – Choluteca

La UNAH, en su oferta académica se encuentran carreras del campo de ciencias, ciencias económicas, administrativas y contables, ciencias espaciales, ciencias jurídicas, ciencias médicas, ciencias sociales, humanidades y artes, ingeniería, odontología, química y farmacia [11]. Estas carreras que se ofrecen a los aspirantes a licenciaturas y especialidades son basadas en estudios de tendencias del empleo que generan los sectores públicos y privados del país. Cabe destacar que la UNAH es la institución de educación superior que más demanda académica genera por la población hondureña y que a finales del 2019, se entregaron más de ocho (8) mil títulos universitarios a profesionales de las diferentes ramas del conocimiento [12].

2.1.2. Organización Jerárquica

La UNAH se estructura de forma jerárquica con diferentes unidades administrativas en cada nivel [13], tal y como se describe a continuación:

- **Dirección Superior:** Consejo Universitario, Junta de Dirección Universitaria.
- **Ejecutivo y Académico:** Rectoría, Abogado General, Secretaría General, Secretaría Ejecutiva de Administración de Proyectos de Infraestructura, Secretaría Ejecutiva de Administración y Finanzas, Secretaría Ejecutiva de Desarrollo Institucional, Dirección Ejecutiva de Gestión y Tecnología, Dirección de Comunicación Estratégica, Vicerrectoría Académica, Vicerrectoría de Orientación y Asuntos Estudiantiles, Vicerrectoría de Relaciones Internacionales, Facultades, Centros Regionales.
- **Control:** Comisión de Control de Gestión, Auditoría Interna.
- **Cuerpo Auxiliar:** Dirección de Educación Superior, Comisionado Universitario.

La Vicerrectoría Académica (VRA), posee diferentes facultades o atribuciones como son: sustituir al Rector en caso de no poder ejercer el cargo, emitir dictámenes, acuerdos y resoluciones, regular el funcionamiento de los programas de educación a distancia y virtual [14], entre otras actividades. La Vicerrectoría Académica, se divide en direcciones académicas; encargadas de dirigir ejes como son: Docencia, Vinculación Universidad – Sociedad, Investigación Científica, Educación a Distancia, Innovación Educativa, Ingreso, Permanencia y Promoción y el Sistema de Admisión a la UNAH [15].

La Dirección del Sistema de Admisión (DSA), se origina gracias a la aprobación del proyecto "Sistema de admisión de los estudiantes del primer ingreso a la UNAH" [3], y es la encargada de asegurar el ingreso transparente, de calidad y sin discriminación de ningún tipo, de los aspirantes, a los títulos de grado, provenientes de centros de educación media o de otras universidades que desean ingresar a la UNAH, así como gestionar el procedimiento de los propios estudiantes que quieran realizar cambio de carrera [16]. En la Figura 1 se describe el organigrama reducido de la unidad de estudios.

Figura 1: Organigrama reducido de la unidad de estudio (Fuente: elaboración propia)

2.1.3. Proceso de admisión

Este nace con la instauración de las Normas Académicas en el año 1992, con el objetivo de mejorar la calidad y rendimiento estudiantil en la educación superior, esto

conociendo el nivel de formación con el que se matriculan los estudiantes universitarios. En el año 2004 el Consejo Universitario aprueba la aplicación de una Prueba de Aptitud Académica (PAA) sostenido por los índices bajos de aprobación que obtienen los estudiantes universitarios en sus asignaturas para ese entonces [3].

Este proceso ha pasado por varias etapas de análisis, validación y seguimiento. Sin embargo, hasta la actualidad se ha mantenido el mismo criterio sobre el puntaje mínimo requerido para ser admitido en la UNAH, tal y como se explicará en la sección 2.1.4. Es importante mencionar que cada carrera universitaria tiene sus propios criterios de admisión, por lo que aprobar la PAA no implica acceder a cualquier carrera.

La primera aplicación de la PAA se realiza a finales del año 2006 y se mantiene hasta la actualidad con ciertas modificaciones (incorporación de nuevas pruebas y criterios de admisión). Sin embargo, mediante ésta la UNAH mide los conocimientos necesarios de los aspirantes que se someten para optar a una titulación de grado o postgrado en modalidad presencial o a distancia, y el cual es de manera obligatoria, según los Artículos 195 y 196 de las Normas Académicas [17], para las siguientes clasificaciones de aspirantes [18]:

- Aspirantes que entrarán a la UNAH después de cursar la educación media.
- Aspirantes para cambio de carrera, los cuales ya están matriculados en la UNAH.
- Aspirantes graduados de la UNAH o de otra universidad nacional o extranjera.
- Aspirantes no graduados de otras universidades nacionales o extranjeras.

En el año 2009, el Consejo Universitario aprueba que los aspirantes puedan realizar la PAA un máximo de tres veces, ya sea por no haber obtenido el puntaje mínimo necesario para ser admitido en la UNAH o por no tener el puntaje requerido para pertenecer a una carrera universitaria de su agrado. En el año 2012, se agrega como requisito para ingresar a la carrera de Medicina, realizar la Prueba de Conocimientos para la Ciencias de la Salud (PCCNS), la cual se ha generalizado para otras carreras en el área de la salud y en el año 2018, se agrega como requisito para ingresar a las carreras de ingeniería, exceptuando Ingeniería Agroindustrial e Ingeniería Acuícola y Recursos Marinos Costeros, aplicar a la Prueba de Aprovechamiento Matemático (PAM) [3].

Desde sus inicios, la DSA ha considerado que el uso de tecnologías informáticas es indispensable para desarrollar de manera eficiente este proceso, por ello, desde la primera aplicación, la inscripción se realizó en línea y a nivel nacional. Esto dio cabida a crear la página web de admisión, donde los aspirantes pueden inscribirse y conocer los resultados de

la prueba, además de informar sobre fechas de aplicación de los procesos, información de interés para aspirantes con discapacidades y mediante la asociación con la Vicerrectoría de Orientación y Asuntos Estudiantiles (VOAE) parte del Nivel Ejecutivo y Académico, conocer otros aspectos como ofertas de estudio, planes de estudios y puntuaciones exigibles en los resultados [3].

Etapas del proceso de admisión

Se observa en la **Error! Reference source not found.** las etapas del proceso de admisión que siguen los aspirantes a realizar estudios universitarios en la Universidad Nacional Autónoma de Honduras.

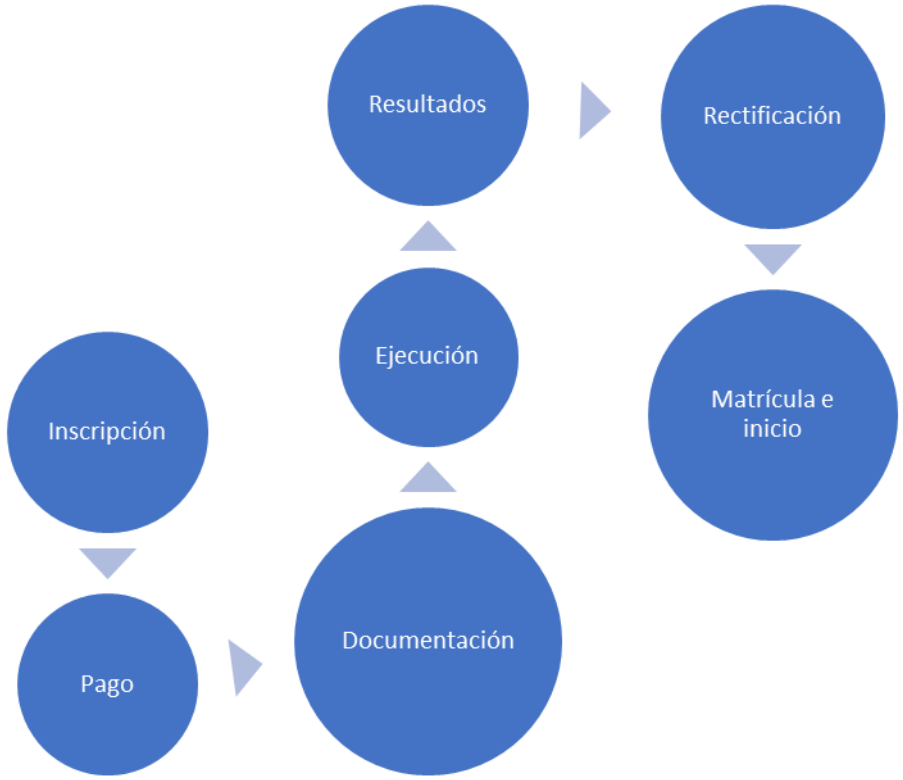

Figura 2 - Etapas del proceso de admisión (Fuente: elaboración propia)

1. **Inscripción**: todo aspirante a la UNAH, iniciará llenando de manera personal el formulario de inscripción proporcionado por la DSA. Se solicitan datos personales, de estudios previos, además deberá elegir tres (3) carreras universitarias para estudiar en orden de prioridad que el aspirante considere conveniente. Al finalizar el formulario el

aspirante recibirá un número de solicitud de admisión que servirá para realizar etapas posteriores.

2. **Pago**: con el número de solicitud generado, se procede a realizar el pago en alguna de las instituciones financieras. Se cobrará un monto adicional si el aspirante por requisito de alguna carrera necesita realizar una prueba adicional (PCCNS o PAM).
3. **Documentación**: una vez el aspirante haya completado el paso 2, deberá presentar un documento de identificación y el recibo de pago realizado de forma presencial en las oficinas de la UNAH, en ese momento al aspirante se le tomará una fotografía y se le entregará una credencial que habilita poder realizar la PAA, además de una guía de estudios donde puede practicar ejemplos de cada parte (verbal y matemática), la estructura de las preguntas y la forma correcta de rellenar el cuestionario de la PAA.
4. **Ejecución**: el aspirante procede a ejecutar la PAA, éste previamente conoce la estructura del cuestionario a rellenar las cuales se explican en la sección 2.1.4.
5. **Resultados**: en un periodo no superior a un mes, los aspirantes pueden visualizar en la página web de la DSA los resultados de admisión. En caso de ser admitido a la UNAH, también podrán ver si fueron admitidos a la primera, segunda o tercera opción que eligieron al momento de la inscripción.
6. **Rectificación**: en esta etapa se realizará, en el caso de ser necesario, cualquier cambio o rectificación de problemas antes de matricularse, por ejemplo, ser admitido a la UNAH, pero no a alguna de las tres opciones de estudios en la inscripción.
7. **Matrícula e inicio**: si el aspirante ha sido **admitido** a la UNAH y ha rectificado los posibles errores como resultados de la PAA, si es que los hubiese tenido, podrá iniciar el trámite de matrícula y las clases en la fecha que le corresponda.

2.1.4. Prueba de Aptitud Académica (PAA)

La PAA es una evaluación de las habilidades y conocimientos previos necesarios que debe tener todo aspirante a estudios universitarios, desarrollada por *College Board* [19]. Su objetivo es ayudar a potenciar, seleccionar y ubicar estudiantes en instituciones de educación superior. La PAA originalmente tiene tres (3) componentes: lectura y redacción, matemáticas e inglés. Sin embargo, la UNAH utiliza solo los primeros dos, dado que para el año 2017 a pesar de impartirse clases de inglés desde la educación primaria en el país, solamente medio millón de habitantes hablan inglés [20]. Además, el puntaje del

componente de inglés no se incluye para realizar el cálculo total, por ello este componente no aparece en el cuestionario que el aspirante rellena.

A continuación, se describe lo que mide cada componente utilizado:

- **Razonamiento verbal (lectura y redacción):** capacidad para comprender, razonar, analizar e interpretar textos, mediante la lectura y análisis de ideas fundamentales.
- **Razonamiento matemático (matemáticas)**: habilidades para procesar, analizar y aplicar la aritmética, el álgebra, la geometría y la estadística.

Ambos componentes tienen un puntaje escalado mínimo y máximo de 200 a 800 puntos respectivamente. La escala se calcula respecto al número de respuestas correctas en el componente matemático y respecto al resultado de cada sección en el componente verbal [21]. La combinación de ambos puntajes permite evaluar los conocimientos necesarios para la admisión de los aspirantes a la UNAH [3].

En la Tabla 2 se muestra la clasificación de los resultados de la PAA según lo plantea College Board [19].

Tabla 2 - Clasificación de los resultados de la PAA

Puntaje PAA Verbal	Puntaje PAA Matemática	Puntaje PAA Combinada	Clasificación
200 a 299	200 a 299	400 a 599	Muy bajas
300 a 399	300 a 399	600 a 799	Bajas
400 a 599	400 a 599	800 a 1199	Promedio
600 a 699	600 a 699	1200 a 1399	Alta
700 a 800	700 a 800	1400 a 1600	Muy alta

Para ser admitido en la UNAH, el aspirante debe obtener un resultado mínimo de 700 puntos [3]. Sin embargo, esto no implica ser admitido en una de las carreras universitarias seleccionadas en la etapa de inscripción.

Para ser admitido en una carrera deberá obtener el puntaje mínimo necesario solicitado por esa carrera, así como realizar una prueba adicional para competir por una plaza si es que ésta lo solicita, tal y como se muestra en el Anexo C – Oferta académica.

2.2. Análisis de datos

Es un proceso que conlleva identificar la correlación, transformación, limpieza y modelado de los datos, con el objetivo de descubrir patrones o información útil que hasta ese momento se encuentra oculta y que es relevante para la toma de decisiones [22].

Las empresas en estos últimos años están teniendo un ambiente muy competitivo, dado que cada vez intentan generar mejores estrategias para incorporarse o mantenerse en el mercado. Las universidades no son una excepción, por ello a través de los años sistematizan más sus procesos y obtienen datos que les permita proporcionar un servicio de calidad.

En tal caso, se suelen utilizar dos tipos de análisis de datos dependiendo de las necesidades operativas tal y como se explican a continuación:

Análisis descriptivo

Este tipo de análisis tiene como objetivo recolectar, organizar, resumir, describir y presentar los datos que corresponden a un conjunto de elementos, esto implica utilizar métodos de distribución de frecuencias para los datos, ya sean cuantitativos; numéricos o cualitativos; subjetivos y no numéricos [23].

La visualización es el método mayormente utilizado para describir los datos con que se cuenta. Dependiendo del tipo de dato (cualitativo o cuantitativo) se deberá realizar la representación visual correspondiente. A continuación, se describen los tipos de datos mencionados con anterioridad:

Cualitativos

Se refieren a cualidades o modalidades que no se expresan numéricamente, como ser: ordinales; son aquellos datos que siguen un orden o secuencia (p. ej. las letras del abecedario, los meses del año) y categóricos; representa aquellos datos que no tienen un orden de prioridad (p. ej. el estado civil) [24]. Estos tipos de datos pueden ser representados mediante gráficos de barras y sectores.

Cuantitativos

Al contrario de los datos cualitativos, estos son representados por cantidades o valores numéricos, como ser: discretos; aquellos datos que toman valores enteros (p. ej. el número de hijos o de alumnos en clases), y continuos; los datos que pueden tomar un valor

dentro de un intervalo real (p. ej. la estatura o peso de una persona) [24]. Este tipo de datos pueden ser representados mediante gráficos de sectores, líneas, dispersión, histogramas, diagrama de Pareto y diagramas de cajas y bigotes.

En la Figura 3 se muestran los componentes del diagrama de cajas y bigotes, el cual está compuesto de un rango sin datos atípicos, datos atípicos, rango intercuartílico (RIC), cuartiles Q1, Q2 y Q3, mediana (Q2), máximo y mínimo [25]. El cuál muestra una representación visual de la media, los cuartiles y el conjunto de datos [26]. Los valores atípicos proporcionan ruido o distorsión al estudio, por tanto, es necesario identificarlos para realizar un tratamiento sobre ellos.

Se consideran valores atípicos (*outliers*) mediante la Ecuación 1, aquellos que salen de los bigotes de la caja que se extiende hasta los valores mínimo y máximo, es decir hasta 1.5 veces el RIC $(Q_3 - Q_1)$. Además, se pueden considerar valores extremadamente atípicos mediante la Ecuación 2, aquellos que exceden 3 veces el RIC.

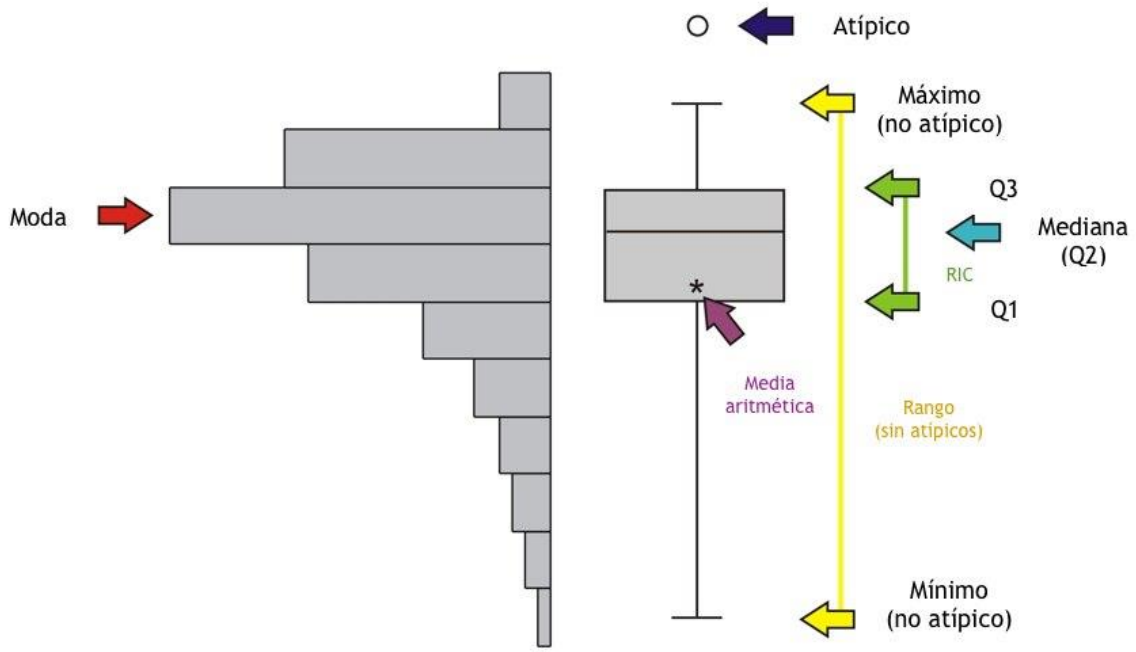

Figura 3 - Componentes del diagrama de cajas y bigotes

$$x\ es\ valor\ atípico \leftrightarrow \begin{cases} x > Q_3 + 1.5(Q_3 - Q_1) \\ x > Q_1 - 1.5(Q_3 - Q_1) \end{cases}$$

Ecuación 1 - Determinación de valores atípicos mediante diagrama de cajas y bigotes

$$x\ es\ valor\ atípico\ \leftrightarrow \begin{cases} x > Q_3 + 3(Q_3 - Q_1) \\ x > Q_1 - 3(Q_3 - Q_1) \end{cases}$$

Ecuación 2 - Determinación de valores extremadamente atípicos mediante diagrama de cajas y bigotes

Generalmente el análisis descriptivo es lo primero que se realiza con un conjunto de datos, dado que estos a primera vista no proporciona información importante, pero que al tratarlos expresan eventos pasados o sucesos históricos que permiten tomar acciones que se adapten a los objetivos de la empresa [27]. Generalmente las organizaciones que utilizan este tipo de análisis poseen un panel de control o *dashboard* donde pueden observar y filtrar rápidamente los datos y centralizar los indicadores claves (KPI, por sus siglas en inglés) que necesitan para saber que está pasando con algún proceso en específico dentro de la empresa.

Análisis predictivo

Los análisis predictivos, consisten en extraer modelos basados en el análisis de datos históricos, que predigan el comportamiento futuro o sean capaz de estimar resultados que no sean fácilmente visibles. Éste tipo de análisis emplea diversas técnicas estadísticas de modelización, aprendizaje automático y minería de datos con el objetivo de elaborar predicciones de cara al futuro [28].

El gran volumen de datos que últimamente las empresas han estado recolectando a través de los años, ha permitido que técnicas y modelos matemáticos surjan nuevamente con el objetivo de pronosticar con cierta probabilidad lo que puede llegar a suceder. Esto no implica predecir al 100% lo que ocurrirá. Sin embargo, se podrá tener una noción clara sobre cómo reaccionar a los posibles eventos futuros. Además, genera un aumento en la competitividad de las empresas, dado que estas buscan una ventaja a la hora de proporcionar sus productos y servicios [29].

2.3. Aprendizaje automático (*machine learning*)

El aprendizaje automático es la rama de la inteligencia artificial que crea sistemas que aprenden de forma automática [30]. En este contexto, aprender significa identificar

patrones complejos en grandes cantidades de datos, donde se aplican algoritmos que analizan los datos y es capaz de pronosticar sucesos futuros. Además, el término automático representa que estos modelos serán capaces de mejorar sin influencia humana, por lo que serán autosostenibles. El *machine learning* es una disciplina enfocada en dos preguntas interrelacionadas "¿Cómo se pueden construir sistemas informáticos que mejoren automáticamente a través de la experiencia?" y "¿Cuáles son las leyes teóricas fundamentales que gobiernan todo sistema de aprendizaje, independientemente de si se implementa en computadoras, humanos u organizaciones?" [31]. El aprendizaje automático cubre diversas aplicaciones desde la clasificación de correo electrónico como spam, hasta aprender a identificar rostros en imágenes y controlar robots para lograr objetivos específicos.

Cuando la cantidad de datos que se recolectan son demasiados para que una sola persona los analice y pueda sacar conclusiones o mejor aún hacer pronósticos, es preferible utilizar algoritmos que detecten automáticamente este comportamiento, interrelacione variables y realice las predicciones. Estos algoritmos generan modelos que son usados para explicar o predecir la realidad [32].

Los modelos son generados mediante el uso de variables de información que es observable en el entorno que lo rodea, por lo que son llamados paramétricos, donde existe un vínculo entre variables independientes, representadas por factores o causas, y dependientes el cual representa el resultado cuya variación se está estudiando. En la Ecuación 3, se aprecia un ejemplo de modelo lineal donde se considera que todos los factores o causas que influyen en la variable dependiente $\boldsymbol{Y}$, se pueden separar en dos grupos, el primero contiene una variable independiente (observable o conocida) $\boldsymbol{X}$ y el segundo se refiere relaciona factores que no pueden ser identificados bajo observación; conocido como error **aleatorio** $\boldsymbol{\varepsilon}$ [32].

$$\boldsymbol{Y} = \beta 0 + \beta 1 \boldsymbol{X} + \varepsilon$$

Ecuación 3 - Ejemplo de modelo lineal [22]

Existen casos donde no es posible determinar cómo se comportan los datos, es decir, no se puede conocer el resultado de una variable dependiente a través del comportamiento

de variables predictoras o independientes [32]. Es ahí cuando se da un protagonismo al aprendizaje automático, donde se toman los datos observados para introducirlos en una "máquina" a partir de los cuales se le enseña a interpretarlos para obtener el modelo.

2.3.1. Técnicas de aprendizaje automático (*machine learning*)

En general, los problemas que se pueden resolver utilizando aprendizaje automático (*machine learning*), son básicamente de dos clases: aprendizaje supervisado (supervised learning) y aprendizaje no supervisado (unsupervised learning) [33]. En la Figura 4 se observa un ejemplo de las técnicas de aprendizaje automático.

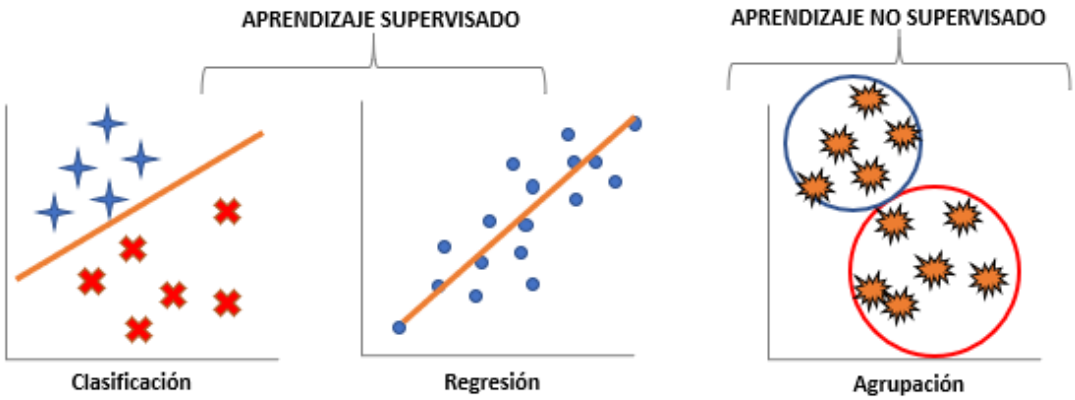

Figura 4 - Técnicas de aprendizaje automático (machine learning) (Fuente: elaboración propia) [21]

Aprendizaje supervisado (Supervised learning)

Dentro de esta categoría se encuentran aquellos problemas cuyo objetivo es pronosticar una o varias de las variables del conjunto de datos. Los algoritmos de aprendizaje para determinar el comportamiento de los datos proporcionados realizan comparaciones entre lo observado y lo predicho, al hacerlo tantas veces como les sea posible se dará cuenta que las predicciones coinciden en mayor o menor parte con las observaciones [34]. El algoritmo tomará de este entrenamiento el modelo con el menor error aleatorio posible y que represente mejor los datos.

Dependiendo de lo que se quiera pronosticar, se tienen dos tipos de algoritmos a utilizar, los cuales se describen a continuación:

Clasificación

Los algoritmos de clasificación utilizan métodos matemáticos para identificar el posible resultado de una variable discreta, en otras palabras, son útiles cuando la respuesta del problema a resolver cae dentro de un conjunto finito de posibilidades. Dependiendo del número de respuestas a considerar, en caso de que el modelo entrenado sea para predecir dos posibles clases objetivos (verdadero o falso) se le conoce como **clasificación binaria**, si se deben predecir más de dos clases se le conoce como **clasificación multicategórica o multiclase** [35].

Regresión

Los algoritmos de regresión intentan predecir un resultado sobre una variable continua, es decir, la respuesta se presenta mediante una cantidad que se determina en función de las entradas del modelo en lugar de limitarse a un conjunto de etiquetas posibles. La regresión lineal, es un ejemplo típico de este tipo de algoritmo, es un método versátil utilizado para predecir precios, análisis de series de tiempo o financieros [35].

Aprendizaje no supervisado (Unsupervised learning)

A diferencia del aprendizaje supervisado, esta categoría intenta identificar estructuras en los datos ya que no se tiene un objetivo específico a predecir, sino que los valores se agrupan según relaciones entre ellos y determinados criterios.

Las principales aplicaciones de aprendizaje no supervisado son:

- Segmentación de conjuntos de datos por atributos compartidos.
- Detección de anomalías que no encajan en ningún grupo, donde a partir de los datos de entrenamiento el algoritmo logra identificar cuáles son los valores “normales” de ciertos parámetros.
- Simplificación del conjunto de datos agregando variables con atributos similares.

En general, su objetivo es identificar la estructura intrínseca de los datos, para agrupar o reducir la dimensionalidad.

2.3.2. Evaluación de modelos de aprendizaje automático (*machine learning*)

Una vez creados modelos de aprendizaje automático mediante el uso de algoritmos de aprendizaje supervisados y no supervisados se deben evaluar para determinar si realizará un buen trabajo con datos futuros. Las nuevas incidencias que se presenten tienen valores con resultados desconocidos, se debe comprobar la métrica de precisión del modelo en relación con los datos que ya conocen la respuesta [36]. Para evaluar correctamente un modelo, se debe separar el conjunto de datos en una muestra representativa que pueda ser utilizada para entrenar el modelo, generalmente es del 70% de los datos originales. Sin embargo, puede variar dependiente del volumen de los datos, y un conjunto de validación con un el 30% restante que permita comparar las predicciones generadas por el algoritmo de aprendizaje con los datos reales que actualmente se conocen.

Se deberán utilizar las métricas de evaluación correspondientes dependiendo del tipo de técnica de aprendizaje utilizada. En la Tabla 3 se muestran las métricas comúnmente utilizadas para cada tipo de modelo de aprendizaje.

Tabla 3 - Métricas de evaluación de modelos de aprendizaje automático (machine learning)

Clasificación	Regresión	Agrupamiento
Matriz de confusión	Error cuadrático medio (MSE)	Similitud global [37]
Precisión (Accuracy)	Error absoluto medio (MAE)	Fmeasure [37]
Recall, exhaustividad o sensibilidad (Tasa positiva real)	Coeficiente de determinación (R^2)	Jaccard-index [37]
Specifity, especificidad (Tasa negativa real)	Error porcentual absoluto medio (MAPE)	
F1-Score		
Receiver Operating Characteristic) (ROC), Área bajo la curva ROC (AUC)		

A continuación, se describirán las métricas de rendimiento para los modelos de aprendizaje supervisado de clasificación y regresión:

Métricas de rendimiento modelos de clasificación

Matriz de confusión

La matriz de confusión es una tabla que describe el rendimiento de un modelo de clasificación en los datos de prueba o validación, donde se desconocen los verdaderos valores de predicción que ha dado como resultado el modelo [38], pero que mediante esta métrica es de fácil evaluación, tal y como se muestra en la Tabla 4.

Las abreviaturas planteadas en la Tabla 4 [39] que se describen a continuación permiten leer la información que proporciona la matriz de confusión:

- **TP** (verdaderos positivos): representa el número de aciertos positivos de la clase real identificados correctamente por el modelo.
- **TN** (verdaderos negativos): es el número de aciertos negativos de la clase real identificados correctamente por el modelo.
- **FP** (falsos positivos): es el número de resultados positivos pronosticados por el modelo cuando debieron ser negativos.
- **FN** (falsos negativos): es el número de resultados negativos pronosticados por el modelo cuando debieron ser positivos.

En un escenario ideal, se espera que la matriz de confusión genere cero (0) falsos positivos y cero (0) falsos negativos, pero esto dependerá de la calidad de los datos y los ajustes realizados al modelo utilizado. Además, dependiendo del problema, será necesario la minimización de falsos positivos o negativos.

Tabla 4 - Matriz de confusión

		Predicción		**Total, realidad**
		Positivo	**Negativo**	
Realidad	**Positivo**	TP	FN	TP + FN
	Negativo	FP	TN	FP + TN
Total, predicción		TP + FP	FN + TN	

Precisión (accuracy)

La precisión o accuracy, representa el porcentaje de elementos clasificados correctamente en comparación al número total de registros. La Ecuación 4, expresa la fórmula correspondiente para su cálculo [38].

Es la medida directa que determina la calidad de un clasificador, comprendida entre valores de 0 a 1, cuanto más alto el valor resultante mejor. Esta métrica no funciona bien en conjuntos de datos donde existe una clase que posee mayor número de elementos que otra (desbalanceo de clases) [39].

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Ecuación 4 - Cálculo de la precisión o accuracy en modelos de clasificación

Recall, exhaustividad o sensibilidad

La sensibilidad es un porcentaje que expresa el número de elementos identificados correctamente como positivos del total de elementos positivos reales [38]. En la Ecuación 5 se expresa la fórmula utilizada para el cálculo del Recall.

$$recall = \frac{TP}{TP + FN}$$

Ecuación 5 - Cálculo del Recall

Precisión de la sensibilidad

La precisión de la sensibilidad es el número de elementos identificados correctamente como positivos del total de elementos positivos pronosticados [38][39]. En la Ecuación 6 se expresa la fórmula utilizada para el cálculo de la precisión de la sensibilidad.

$$precision = \frac{TP}{TP + FP}$$

Ecuación 6 - Cálculo de la precisión de la sensibilidad

El recall describe un modelo de clasificación con respecto a los falsos positivos (número de elementos erróneamente clasificados del total real), mientras que la precisión

describe el rendimiento del modelo con respecto a los falsos positivos (número de elementos correctamente acertados del total pronosticado) [38]. Si la intensión es disminuir el número de falsos negativos se deberá tratar de que el Recall sea lo más cercano al 100%, de lo contrario, si se pretende disminuir el número de falsos positivos se deberá tener una alta precisión.

Especificidad

La especificidad representa el porcentaje de elementos correctamente identificados como negativos del total de elementos negativos reales [38] tal y como se expresa en la fórmula planteada de la Ecuación 7.

$$specificity = \frac{TN}{TN + FP}$$

Ecuación 7 - Cálculo de la especificidad

F1-Score

Es indispensable establecer una métrica de evaluación de un solo número para su posterior optimización. La precisión es un ejemplo de métrica de evaluación de un solo número que permite comparar rápidamente dos clasificadores, mientras que la sensibilidad y especificidad no permite la evaluación esta evaluación directa dado que genera dos resultados en el caso de la sensibilidad (recall y precisión) [38]. F1-Score plantea combinar las métricas de precisión y recall tal y como se muestra en la Ecuación 8.

$$F1 = \ 2 * \frac{precision * recall}{precision + recall}$$

Ecuación 8 - Cálculo de F1-Score

Área bajo la curva ROC

El área bajo la curva ROC es una métrica popular utilizada por la industria, esto porque la curva es independiente del cambio en la proporción de la respuesta [40]. Representa las habilidades del modelo para discriminar entre clases positivas y negativas. Un área de 1, representa que el modelo realiza las predicciones perfectamente, un área de

0.5 representa que el modelo es bueno o aleatorio [41]. En la Figura 5 se muestra un ejemplo de la curva ROC (AUC) donde existe una alta relación entre los valores correctamente identificados como positivos (sensibilidad) y correctamente identificados como negativos (especificidad) en comparación al conjunto de datos total.

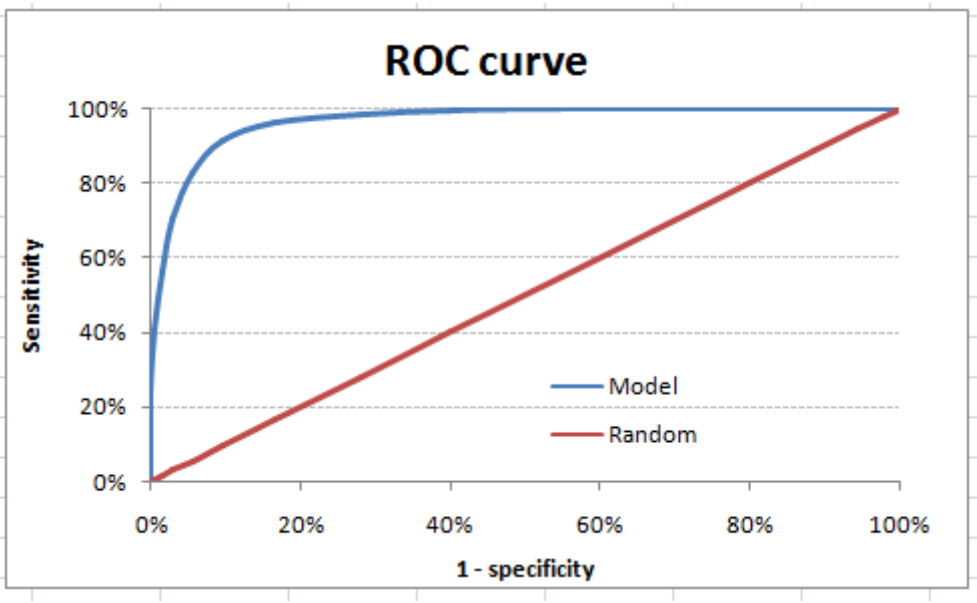

Figura 5 - Ejemplo de curva ROC (AUC) [33]

Métricas de rendimiento modelos de regresión

Error cuadrático medio (MSE)

El MSE básicamente, mide la diferencia cuadra entre las predicciones y el valor real dividida por el total de elementos en evaluación [42], tal y como se expresa en la Ecuación 9, donde y_i es el resultado real esperado y y_i' es la predicción del modelo.

Cuanto menor sea el MSE, más precisas serán las predicciones del modelo de regresión, donde un MSE de 1 implica que el modelo predice perfectamente.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2$$

Ecuación 9 - Cálculo del error cuadrático medio (MSE)

Error absoluto medio (MAE)

El error absoluto medio (MAE), es una puntuación lineal donde todas las diferencias individuales se ponderan por igual en el promedio [42], tal y como se plantea en la Ecuación 10. Es importante saber que esta métrica no es tan sensible a valores atípicos como el error cuadrático medio.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|$$

Ecuación 10 - Cálculo del error absoluto medio (MAE)

Coeficiente de determinación (R^2)

No es tan sencillo identificar si el modelo generado es bueno o no con conocer el valor de MSE, por lo que es necesario determinar otra métrica que apoye a este análisis. El coeficiente de determinación (R^2) está relacionada con el MSE, pero en una ponderación porcentual de $-\infty$ a 1 [42]. En la Ecuación 11, se muestra la fórmula para el cálculo del coeficiente de determinación donde MSE (model) es el mismo comentado en la Ecuación 9, mientras que MSE (baseline) realiza la diferencia entre el valor real (y_i) y la media de la observación (y_i^*) [42].

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)},$$

$$MSE(baseline) = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)^2$$

Ecuación 11 - Cálculo del coeficiente de determinación

Error porcentual absoluto medio (MAPE)

El error porcentual absoluto medio es un indicador que mide el tamaño del error (absoluto) en términos porcentuales. Es de fácil interpretación y utilización debido a su magnitud porcentual, en la Ecuación 12, se muestra la fórmula utilizada para el cálculo del MAPE.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - y_i'}{y_i}\right|$$

Ecuación 12 - Cálculo del error porcentual absoluto medio (MAPE)

2.3.3. Cuadro de mando (*Dashboard*)

El cuadro de mando es una herramienta de gestión que permite convertir la estrategia de la organización en objetivos operativos, de esta forma, permite realizar un seguimiento de las acciones encaminadas para alcanzar la visión de la organizacional [43]. El cuadro de mando puede ser implementado a nivel operativo o departamentos con visión y estrategias de negocios definidas.

A continuación, se describen las cuatro (4) perspectivas a las que se alinean los objetivos operativos e indicadores [44]:

- **Financiera**: vincula los objetivos de cada unidad del negocio con la estrategia global de la empresa.
- **Del cliente**: determina los segmentos de clientes y mercado donde se competirá, por tanto, evalúa las necesidades de los clientes con el fin de alinear los productos y servicios con sus preferencias.
- **Procesos internos**: define la cadena de valor de los productos con el fin de entregar a los clientes soluciones a sus necesidades.
- **De aprendizaje y crecimiento**: se obtiene la información necesaria para lograr los resultados de las perspectivas anteriores.

Para su implementación, primeramente, se debe diseñar un mapa estratégico de la organización o departamento, tal que para cada perspectiva se determinen el conjunto de objetivos que son realmente importantes para alcanzar la visión planteada [43]. Cada conjunto de objetivos según la perspectiva está estrechamente relacionado entre sí por causa – efecto, esto implica que alcanzar uno de ellos permitirá acercarnos a otros objetivos de otras perspectivas.

Capítulo 3

3. Metodología

3.1. Descripción del problema

En Honduras, para el año 2017, solo el 16% de los egresados de educación media ingresan a una universidad pública o privada [45]. En cuanto a la UNAH existen grandes dificultades respecto al nivel de conocimientos que los aspirantes deben tener previo a someterse a la PAA, esto puede ser debido a diferentes factores involucrados en su formación, entre estos factores están, la ubicación geográfica de residencia o nacimiento con respecto al centro de estudios de educación media, sexo, carrera de estudios medios, o bien por los bajos ingresos que posee la población hondureña.

Para afrontar este problema, la Dirección del Sistema de Admisión (DSA) de la UNAH en cada uno de los centros regionales, al final de cada proceso, reúne a los encargados de los centros de educación media para mostrar cuales fueron los resultados de los aspirantes provenientes de esos sitios, esto les permite generar estrategias para que sus aspirantes cuenten con los conocimientos necesarios previo a someterse a la PAA, y así éstos puedan obtener mejores puntajes y optar a las carreras universitarias de su agrado.

Este proceso se ha realizado hasta la fecha de manera casi manual donde, el encargado de la DSA en cada centro regional solicita a la central en Ciudad Universitaria la información pertinente a su región para generar sus propios resúmenes. Esto generalmente son análisis descriptivos donde centralizan la información relevante para los encargados de los centros de educación media. Sin embargo, considerando toda la información que se recopila de los aspirantes al momento de realizar la inscripción, sería conveniente utilizar tecnologías para generar el mismo tipo de análisis en todas las regiones de manera dinámica,

que permitan indagar con profundidad los factores que afectan a los aspirantes al momento de realizar la PAA.

El volumen, velocidad y variedad de los datos que se recolectan en cada proceso, no se ajusta directamente a un proyecto Big Data. Sin embargo, es posible estudiar la veracidad con la que completa el formulario de inscripción y encontrar un valor añadido a lo que actualmente se realiza, el cual puede proporcionar oportunidades para definir estrategias que apoyen a los aspirantes al momento de realizar la PAA.

3.2. Metodología

La competitividad de toda empresa se mide en la obtención rápida y eficiente de la información sobre sus procesos [46]. De esta manera, también las tecnologías que son piezas útiles y necesarias para el mejoramiento de estos procesos han tenido un crecimiento acelerado. Una de estas herramientas es la minería de datos, la cual es parte de un proceso conocido como KDD (Discovery Knowledge in Databases), es un proceso organizado alrededor de cinco (5) fases:

1. Integración y recopilación.
2. Preparación de los datos.
3. Minería.
4. Evaluación.
5. Difusión y uso de modelos.

La minería de datos proporciona un conocimiento que permite definir asociaciones, patrones o reglas que en un principio estarán ocultas o desconocidas para los encargados de revisar los datos, pero que permitirá realizar modelos descriptivos o predictivos según la necesidad del negocio.

En este trabajo, la metodología utilizada basada en minería de datos será una combinación entre CRISP-DM (Cross-Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, Assess). Cada una de las metodologías anteriores se enfocan en el descubrimiento de patrones como herramientas de apoyo al negocio y al desarrollo óptimo de proyectos de análisis de datos [46].

En la Figura 6 [46] se describe la estructura del ciclo de vida de la metodología CRISP-DM, conformada por tareas que interactúan entre sí con el fin de determinar el desarrollo óptimo del proyecto.

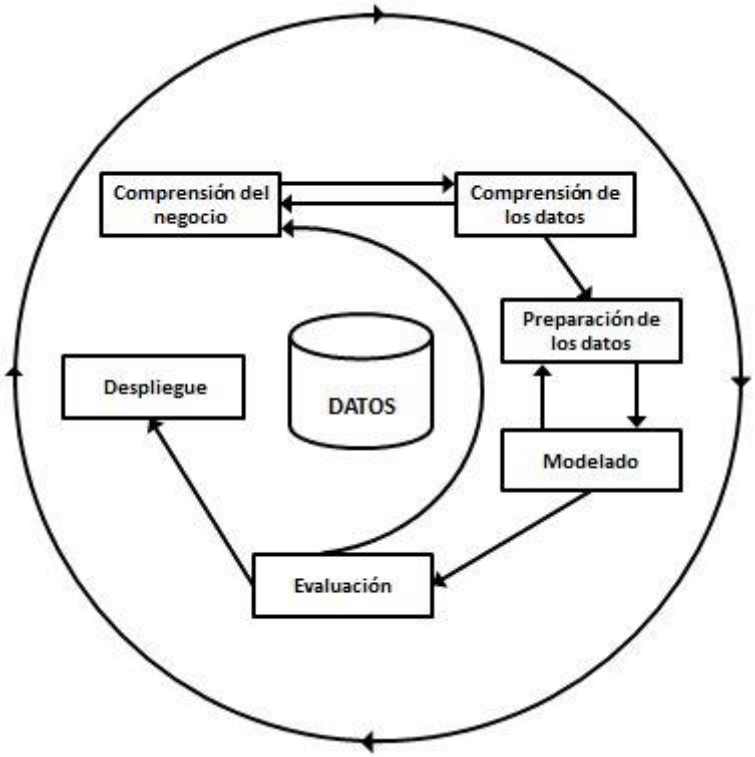

Figura 6 - Fases de la metodología CRISP-DM (Fuente: elaboración propia)

A continuación, se describen las fases de la metodología CRIPS-DM:

1. **Comprensión del negocio**: esta fase se enfoca en el conocimiento de la organización y los objetivos del negocio, por ello es imprescindible contar con la colaboración del personal de la empresa que trabaja en el mismo proceso.
2. **Comprensión de los datos**: consiste en recolectar los datos para describirlos y explorarlos, no solamente con visualizaciones, sino con el uso de técnicas estadísticas que permitan encontrar mayores detalles.
3. **Preparación de los datos**: una vez que se comprenden los datos, se podrá entender el comportamiento de estos. Además, se deben seleccionar y limpiar los datos no pertenecientes al mismo espacio de tiempo para lograr modelos más estables de minería en la etapa de diseño.
4. **Modelado**: ahora, se procede a seleccionar técnicas de modelamiento adecuados a la minería de datos de acuerdo con el contexto del negocio. Además, esta

selección debe estar basada en las métricas de evaluación que caracterizan los modelos, como se comenta en la sección de Análisis de datos.

5. **Evaluación**: es una de las más importantes, dado que permite comprobar la funcionalidad del modelo revisando las etapas anteriores e identificar los próximos pasos.
6. **Despliegue**: en esta fase se revisa el despliegue, documentación, presentación de resultados y mantenimiento de la aplicación para producir un reporte final que englobe todo el proyecto.

En la Figura 7 [47] se muestran las diferentes etapas del dato dentro de la metodología SEMMA, la cual se centra en el desarrollo del proceso de minería de datos y no en los objetivos empresariales. El nombre SEMMA representa la inicial de cada fase (*Sample, Explore, Modify, Model, Assess*).

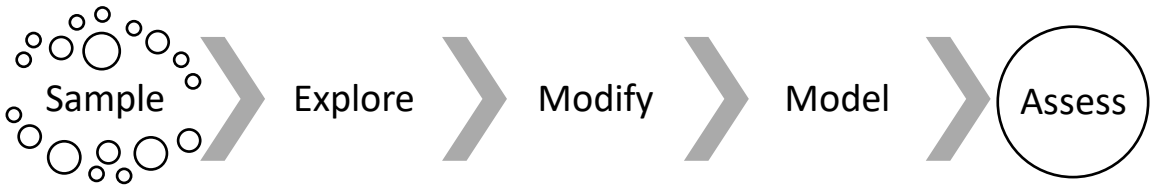

Figura 7 - Fases de la metodología SEMMA (Fuente: elaboración propia)

A continuación, se describen las fases de la metodología SEMMA:

1. ***Sample***: es la primera fase del proyecto. Se preparan los datos que pasaran a exploración [48], previamente, se realiza una partición del conjunto de datos para obtener una muestra representativa y un conjunto de validación, para posteriormente comenzar el análisis de los mismos [46].
2. ***Explore***: consiste en la exploración de los datos a través de técnicas estadísticas que permitan detectar datos anómalos para su posterior tratamiento [46]. Es una de las partes más trabajosas y entretenidas del proyecto [48].
3. ***Modifiy***: en esta fase, se realiza la selección y transformación de los datos [46] la cual permitirá contar con los datos de calidad para la selección y diseño del

modelo final, sobre esta etapa se realiza el tratamiento a los datos anómalos identificados en la fase de exploración (*Explore*).

4. ***Model***: una vez se cuenta con los datos de calidad se procede a la selección del modelo. Se puede considerar cualquier algoritmo de aprendizaje supervisado y no supervisado, así como hacer comparativas entre ellos.
5. ***Assess***: consiste en la evaluación del modelo final. Se utilizarán las métricas correspondientes dependiendo del tipo de modelo seleccionado [46], esto como se comentó en la sección 2.3.2, implicará utilizar el conjunto de validación ya que los datos que proporcione permitirá obtener respuestas en un principio desconocidas por el modelo, y así realizar la comparación.

Se considerará una combinación de ambas debido a que SEMMA se enfoca en el desarrollo del tratamiento y modelado de los datos, en donde aporta la separación de los datos en muestra representativa y validación. Mientras que CRISP-DM enfoca el tratamiento de estos datos en dirección de las necesidades del negocio. La adaptación no es una novedad, sino que se limita a utilizar las metodologías existentes con el fin particular al desarrollo del proyecto. En la Figura 8 se observa como todas las fases de SEMMA están relacionadas con CRISP-DM, en cambio CRISP-DM relaciona los datos con el negocio y más allá del modelado y validación el trabajo futuro que debe existir en todo proyecto de data mining.

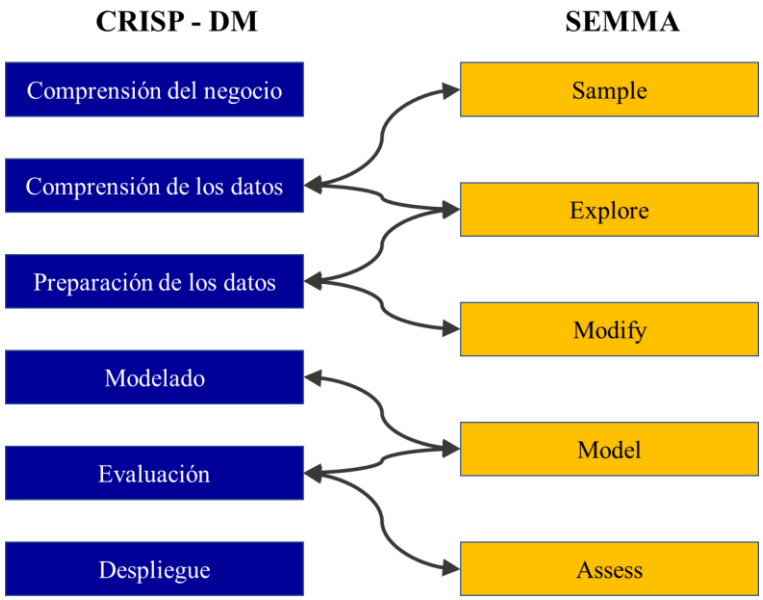

Figura 8 - Comparación metodologías CRISP-DM y SEMMA (Fuente: elaboración propia)

3.3. Planificación del desarrollo del proyecto

La estimación de tiempo necesario para el desarrollo del proyecto se considera en base a las fases de la metodología a seguir. En la Figura 6 se observa el diagrama de Gantt el cuál indica el nombre de cada fase, fecha de inicio y fecha de fin. En general el proyecto se desarrolla en 75 días hábiles comprendidos del 15 de mayo al 27 de agosto del 2020 con una dedicación de 3 horas diarias.

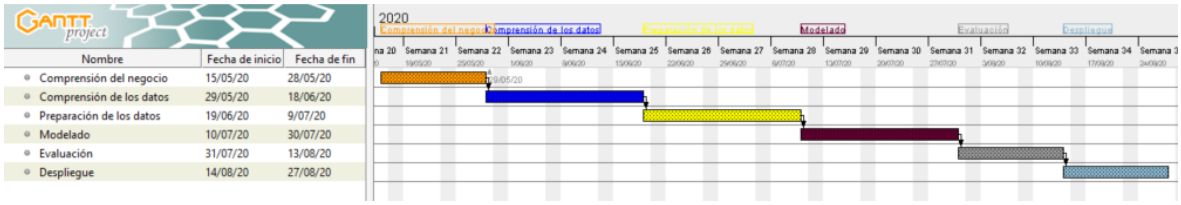

Figura 9 - Fases del trabajo y estimación temporal - Diagrama de Gantt

Las fases comprensión del negocio, evaluación y despliegue requieren de 10 días hábiles cada una para su finalización, mientras que se destinan 15 días hábiles por su nivel de importancia a las fases comprensión de los datos, preparación de los datos y modelado.

3.4. Presupuesto

Para la generación del presupuesto asociado al Trabajo Final de Máster, se tuvieron en consideración las herramientas necesarias de trabajo descritas en el sección 4.1. Además, de la estimación de tiempo expuesto en la sección anterior.

Se describe en la Tabla 5 el presupuesto necesario en componente hardware y servicio de internet para el desarrollo del proyecto.

Tabla 5 - Presupuesto, servicios y componentes hardware

Componente	Vida útil (años)	Coste total
ASUS (GL703GS-E5011)	6	**1,400€**
Servicios de internet	1	**348€**
Total		**1,748€**

En la Tabla 6 se describen los componentes software que se utilizan para el desarrollo del trabajo planteado. Estos componentes son de uso libre, excepto Power BI donde se

considera una licencia BI Pro para un usuario, por lo que el coste es inferior al del componentes hardware.

Tabla 6 - Presupuesto, componentes software

Componente	Vida útil (años)	Coste total
Anaconda	-	-
Python	-	-
XAMPP	-	-
Power BI	1	**119.88€**
Total		**119.88€**

Para esta estimación se ha contado con un único ingeniero informático con salario de 25,500€/año equivalente a 12.60 €/h. El Trabajo Final del Máster se desarrolla en 75 días hábiles estimados en el intervalo comprendido del 15 de mayo al 27 de agosto 2020 con una dedicación de 3 horas diarias, el total devengado por el ingeniero sería el siguiente:

Total-devengado = 75 días * 3 h/días * 12.60 €/h = 2,835 €

El presupuesto general del Trabajo Final del Máster es de **4,702.88 €** calculado sumando los presupuestos hardware, software y salario devengados tal y como se describe a continuación:

Componentes hardware	1,748.00 €
Componentes software	119.88 €
Salario devengado	2,835.00 €
Total	**4,702.88 €**

Capítulo 4

4. Desarrollo de la metodología

Como se comentó en la sección 3.2, se utilizará una combinación entre SEMMA y CRISP-DM, donde la primera metodología está contenida en la segunda. Por ello, se plantean a continuación las fases de CRISP-DM considerando que se desarrollará el tratamiento de los datos orientado al data mining y necesidades del negocio.

4.1. Comprensión del negocio

Comprender los objetivos y requerimientos del proyecto, desde el punto de vista del negocio, es indispensable para transformar esos conocimientos en la definición de un problema de minería de datos y generar un plan preliminar para alcanzar los objetivos.

Determinar los objetivos del negocio

Como se comentó anteriormente, el objetivo de la UNAH es proporcionar a la comunidad en general la posibilidad de optar a un título universitario que cumpla con los estándares de calidad. Sin embargo, para el aspirante, el proceso de admisión resulta ser un tanto complicado debido a los conocimientos que los estudiantes deben poseer para alcanzar el puntaje necesario para ser admitido.

Para la DSA ciertamente es de utilidad identificar los factores que afectan al aspirante para que éste apruebe o no el examen de admisión, así como determinar ese grupo de aspirantes vulnerables a reprobar la PAA y tomar decisiones que permitan mejorar el rendimiento durante el proceso completo. Por ello, se plantearon ciertos requerimientos en

forma de preguntas, tal que permita identificar las características para realizar futuros modelos de aprendizaje.

- ¿Cuál es el sexo de los aspirantes que más se someten a la PAA? y ¿Cuál es su resultado de admisión?
- ¿Cuáles son los estadísticos en la edad de los aspirantes que intentan estudiar en la UNAH?
- ¿Cuál es la carrera universitaria más demandada por los aspirantes de las regiones?, ¿Se encuentra esta carrera en el centro educativo de la región seleccionado?
- ¿Los aspirantes se someten a la PAA en el centro universitario más cercano o en el mismo centro universitario con interés de estudio?
- ¿El factor económico afecta en el resultado de admisión de los aspirantes que realizan la PAA para entrar a la UNAH?
- ¿La familia influye en el resultado de admisión a la UNAH?
- ¿El centro de estudios medios de procedencia es un factor clave para identificar aquellos aspirantes que tienen mayores posibilidades de reprobar la PAA?
- ¿Someterse a la PAA en los años cercanos a la graduación de educación media facilita poseer los conocimientos necesarios para ser admitido en la UNAH?
- ¿Cuál es el porcentaje de aspirantes que realizan más de una vez el examen de admisión?
- ¿Haber sido poseedor de una beca previo a someterse a la PAA implica mayor éxito para ser admitido en la UNAH?

Evaluar el contexto

En la Tabla 7 se muestran los recursos necesarios para determinar la admisión de los aspirantes a la UNAH.

Tabla 7 – Recursos necesarios para el desarrollo del trabajo

Recursos	Datos recopilados de cada proceso de admisión, estos se refieren a todos los que involucren a la Región Sur de Honduras.
	Reglamento de admisión de los estudiantes.
	Normas académicas.
	Estructura del College Board.
Personas	Responsable de la DSA en los centros regionales y CU.

Determinar los objetivos de minería de datos

Una vez conocidos los objetivos del negocio, se procede a convertir éstos en términos más técnicos de minería de datos, así como identificar la forma en que se evaluarán los criterios de éxito.

A continuación, se plantean algunos objetivos de minería de datos en base a las necesidades determinadas del negocio:

- Realizar visualizaciones para detectar valores atípicos en los datos.
- Identificar que aspirante obtendrá el puntaje mínimo necesario como resultado de la PAA para ser admitido en la UNAH.
- Determinar los factores (campos) involucrados en el resultado de admisión que obtienen los aspirante.

Herramientas necesarias de trabajo

Es indispensable identificar una valoración especial del tipo de herramientas y técnicas que se pueden requerir en el trabajo, dado que la selección de estas puede influir en el proyecto completo. Las conversaciones con los encargados de la Dirección del Sistema de Admisión en el Centro Universitario Regional del Litoral Pacífico, así como el Director en Ciudad Universitaria, han colaborado en identificar las herramientas y técnicas que se utilizarán para alcanzar el éxito en los objetivos del proyecto *(comunicación telemática, mayo 2020).* A continuación, se describen las herramientas utilizadas a lo largo del trabajo:

Anaconda

Figura 10 - Logo Anaconda

Es una tecnología construida por científicos de datos para científicos de datos *(Logo,* Figura 10*)*. Esta herramienta contiene varios paquetes y herramientas, como se muestra en la Figura 11, que son utilizados por más de 20 millones de usuarios en todo el mundo, es de código abierto y diseñado para realizar ciencia de datos y aprendizaje automático (*machine learning*) con Python y R.

Proporciona soluciones flexibles para construir, distribuir, instalar, actualizar y administrar software de manera multiplataforma. Hace sencilla la administración de múltiples ambientes de desarrollo las cuales se pueden mantener y ejecutar por separado sin interferencias unas con las otras [49]. Por ello, es ideal para el trabajo a realizar, ya que independientemente de la cantidad de proyectos con el que ya se cuente, las dependencias de librerías necesarias para este no se traslaparán con ningún otro en ejecución.

Figura 11 - Paquetes y herramientas que proporciona anaconda (Fuente: anaconda documentation) [23]

Cabe mencionar que Anaconda es utilizado tanto para hacer aprendizaje automático (*machine learning*) escalable, así como realizar visualizaciones dinámicas para la exploración, generación de modelos y evaluación de los modelos generados.

Python

Figura 12 - Logo Python

Es un lenguaje de programación interpretado, enfocado en la generación de código legible, multiparadigma y multiplataforma *(Logo,* Figura 12*).* Posee licencia de código abierto denominada Python Software Fundation License.

¿Por qué Python y no Scala?

Basado en el informe de resultados combinados de la cuarta encuesta sobre el ecosistema de los desarrolladores llevada a cabo por JetBrains desde inicios del 2020 [50], los tres (3) lenguajes de programación más utilizados en los últimos 12 meses son JavaScript, Java y Python, tal y como se muestra en la Figura 13 a diferencia de Scala que es uno de los menos utilizados y viendo al futuro no es factible.

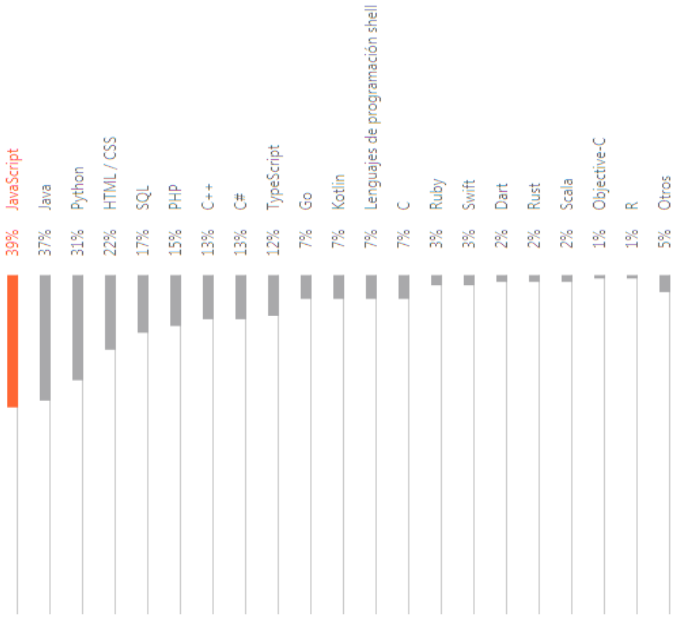

Figura 13 - Lenguajes de programación más utilizados (Fuente: JetBrains) [24]

En cuanto a su rendimiento, a pesar de Scala ser unas 10 veces más rápido que Python [51], está considerado para grandes cantidades de datos. Sin embargo, este trabajo no consiste en el análisis de grandes volúmenes de datos

Jupyter

Figura 14 - Logo Jupyter

Es un proyecto para desarrollo de código abierto y servicios de computación interactiva [52] *(Logo, Figura 14).* Contiene diferentes paquetes de código abierto como Jupyter Notebook, la cuál es una aplicación de para la web que permite crear y compartir documentos que contienen código vivo, ecuaciones, visualizaciones y texto narrativo. Generalmente es utilizado para realizar limpieza y transformación de datos, generación de modelos estadísticos, visualización de datos, aprendizaje automático, entre otros. Por lo que es idóneo su utilización en este trabajo.

XAMPP

Figura 15 - Logo XAMPP

Es una distribución de Apache completamente gratuita y fácil de instalar que contiene MariaDB, PHP y Perl [53] *(Logo, Figura 15).* Los encargados de la DSA en conversaciones comentaron que se utilizaba un SQLServer para almacenar la información recopilada de la inscripción y PAA que realizan los aspirantes como parte del proceso de admisión. Una forma sencilla de simular el entorno de la base de datos es utilizar MySQL dado que es un sistema de gestión de bases de datos relacional, al igual que SQLServer, desarrollado bajo licencia dual: licencia pública general/ licencia comercial por Oracle Corporation [54]. Es aquí donde se almacena los resultados del procesamiento que se realizan en los datos y que se explicará en la sección 4.2 del desarrollo de la metodología.

Power Business Intelligence (Power BI)

Figura 16 - Logo Power BI

Es un conjunto de herramientas de análisis empresarial que permite poner el conocimiento al alcance de todos en la organización. Power BI está integrado en Office 365 [55] *(Logo, Figura 16),* permite conectarse a cientos de orígenes de datos, preparar los datos de forma simplificada y generar análisis ad hoc. Además, permite dinamizar los datos con paneles e informes. Power BI permite concentrar diferentes tipos de datos desde diversos orígenes, además de preparar y modelar los datos con facilidad, eficiente para usuarios conocedores de Excel e intuitivo para quienes lo desconocen. Los encargados de la DSA comentan el interés del uso de esta herramienta, ya que parte de su plan de acción es contar con una licencia que incluya Power BI y permitir que los encargados en los diferentes Centros Regionales puedan tener acceso a esta información de manera rápida y eficiente.

4.2. Comprensión de los datos

Una vez que se comprende las necesidades del negocio, se debe capturar los datos a analizar para familiarizarse con ellos, identificar los posibles problemas de calidad y determinar subconjuntos de datos que puedan validar los análisis posteriores.

Recolección de los datos

La DSA recolecta los datos a través del formulario de inscripción que la aspirante rellena como primera etapa del proceso de admisión. Estos son almacenados en una base de datos SQL Server para su posterior procesamiento tal y como se comentó con anterioridad.

Descripción de los datos

La Dirección del Sistema de Admisión (DSA), a través de conversaciones realizadas por correo electrónico, proporcionó para el análisis tres (3) ficheros en formato ".xlsx" que cubren la aplicación de la PAA en la región sur del país, de diciembre 2006 a septiembre 2019, para que sirvan como datos pilotos y representativos ante el proceso completo de admisión en la UNAH. Cada uno de ellos con el mismo número de campos, y tipos de variables. Uno de los campos proporcionados es el documento de identificación del aspirante previamente anonimizado, por lo que existe un compromiso de responsabilidad en el uso de los datos. Se proporcionan tres ficheros para determinar los siguientes grupos de aspirantes que involucren a la zona sur del país:

- Aspirantes con centro de aplicación CURLP.
- Aspirantes con centro de aplicación distinto al CURLP, pero con centro de estudios CURLP.
- Aspirantes procedentes de los departamentos de Choluteca y Valle con centro de aplicación distinto al CURLP y centro de estudios distinto al CURLP.

Adicionalmente, bajo investigación de los datos públicos que proporciona la Dirección del Sistema de Admisión, se obtuvo un conjunto de datos con la oferta académica que proporciona la UNAH, la cual cuenta con el nombre de la carrera universitaria, facultad, modalidad, puntaje mínimo solicitado, determinación de realizar o no una prueba adicional, así como si tiene que competir o no por una plaza.

Exploración de los datos

A continuación, se explorarán los datos mediante el uso de tablas, gráficas, selección de variables y la aplicación de aprendizaje no supervisado que nos ayuden a agrupar o descartar aquellos campos que no sean relevante con el objetivo planteado.

Del primer proceso efectuado en el año 2006, al último proceso del 2019, en la región sur fueron 33,999 los aspirantes que se han sometido a realizar la Prueba de Aptitud Académica con el objetivo de estudiar en la UNAH. En el Anexo B - Característica de los datos proporcionados, se describen los campos que contienen los ficheros, su tipo de dato y la nueva nomenclatura utilizada y en la Tabla 8 se aprecia un resumen de los 54 campos proporcionados, se identifican campos del tipo fecha, numérico y texto, en los que incluyen información sobre la inscripción y resultados de la PAA, no se incluye información de las matrículas, dado que esta información es tratada por responsables de otra unidad académica. Cabe mencionar que el campo **identity_code** es el documento personal de identificación del aspirante, el cual ha sido anonimizado por la DSA antes de proporcionarlo para este trabajo.

Tabla 8 - Resumen de los datos proporcionados

Tipo de dato	Cantidad
fecha	1
numérico	11
texto	42
Total	**54**

En la Tabla 9 se muestran para cada campo dentro del conjunto de datos el porcentaje y número de registros ausentes o nulos. En este caso, solo se han encontrado valores nulos para los campos birth_year y graduation_year dado que son campos que el aspirante debe rellenados de forma manual.

Tabla 9 - Identificación de valores nulos en los campos

Campos	Registros nulos	Porcentaje
birth_year	79	0.23%
graduation_year	480	1.41%

Dado que estos campos son importantes para realizar un razonamiento que permita explicar los futuros modelos se procedió a descartar estos registros del conjunto de datos

total. Por tanto, los datos proporcionados se reducen al 98.36% equivalentes a 33,443 registros en total.

Como se comentó en la sección 2.3.2, evaluar correctamente un modelo implica separar el conjunto de datos en una muestra representativa para entrenar los modelos generados y un conjunto de validación con datos desconocidos para el modelo que permita comparar las predicciones con los valores reales. Por ello, se procede a dividir de forma aleatoria, mediante la librería ***train_test_split*** perteneciente a ***sklearn***, el conjunto de datos proporcionados en un 85% para muestra representativa y 15% para validación de los futuros modelos generados. El porcentaje destinado para el conjunto de validación es debido al poco volumen de datos y al uso de validación cruzada, sobre el conjunto de muestra, que se realizará con el fin de seleccionar el mejor modelo, tal y como se explicará en la sección 4.4 parte del desarrollo de la metodología. En la Tabla 10 se observa que el número de registros proporcionados limpios de valores ausentes o nulos fueron 33,443, así como la división generada por ***train_test_split.***

Tabla 10 – Número de registros proporcionados divididos para la muestra y validación

	Registros	Campos
Proporcionados/ reducidos	33,443	54
Muestra	28,426	54
Validación	5,017	54

Se observa en la Figura 17 como el número de aspirantes por fecha de aplicación mantiene un comportamiento similar tanto para los datos de muestra como para los datos de validación a pesar de haber sido seleccionados de manera aleatoria. El número de aspirantes que realizan el examen de admisión para entrar a la UNAH ha crecido a través del tiempo, excepto en el tercer proceso 2016 donde este valor para esas fechas decayó drásticamente, esto debido protestas por parte de la comunidad estudiantil que impidieron la realización de procesos administrativos, incluido el de admisión [56]. En la Figura 18 se aprecia que la mayoría de los aspirantes, que involucran la región sur del país y que realizan la PAA para entrar a la UNAH solo la hacen una vez, en esta ocasión identificado por el 79.4% del total de la muestra y el resto de los aspirantes decidió utilizar su segundo o tercer intento, ya sea para obtener el puntaje mínimo requerido para ser admitido o porque el puntaje obtenido le ha permitido ingresar a la UNAH, pero no es suficiente para pertenecer a la carrera universitaria de su agrado.

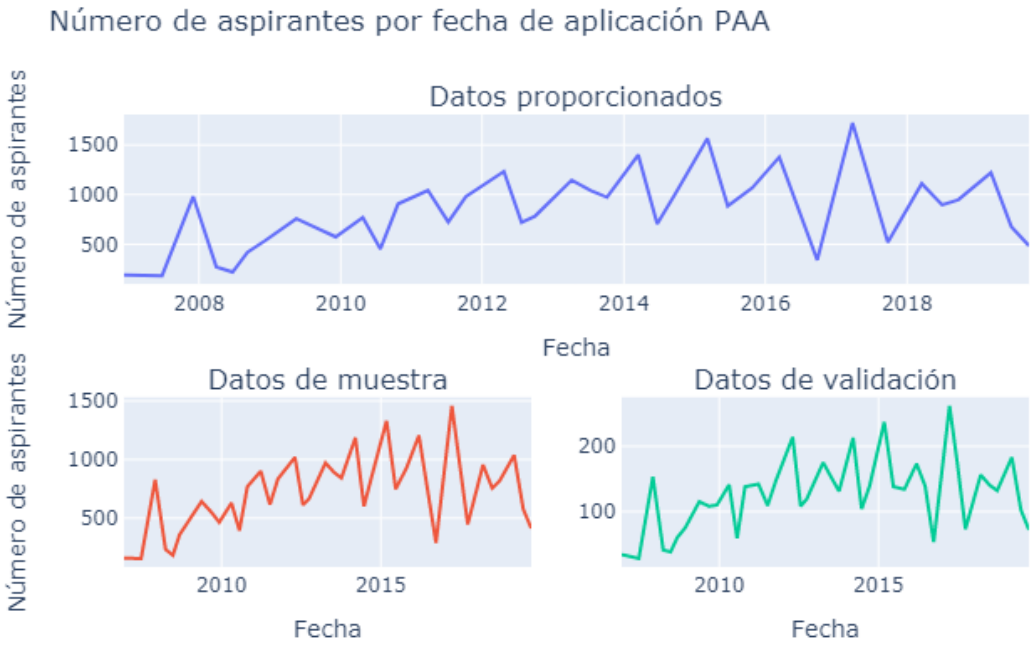

Figura 17 - Número de aspirantes por fecha de aplicación PAA (fuente: elaboración propia)

Figura 18 - Aspirantes según el número de veces que ha realizado la PAA (Fuente: elaboración propia)

La Figura 19 separa de la muestra a los aspirantes que realizaron exactamente dos (2) intentos para ingresar a la UNAH, en su primer intento se observa que el 79.8% de ellos no fueron admitidos, lo que muy probablemente los haya motivado a utilizar su segundo intento, mientras que el 20.2% intentará mejorar su puntaje para ser admitido en alguna carrera específica cuyo puntaje de admisión es muy superior al de la UNAH.

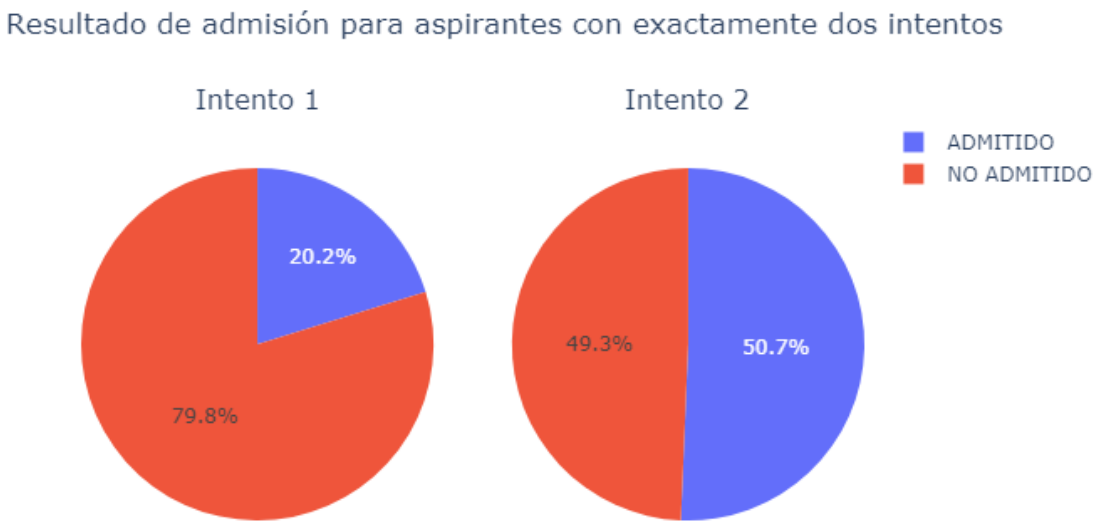

Figura 19 - Resultado de admisión para aspirantes con exactamente dos intentos (Fuente: elaboración propia)

De igual forma, se separa de la muestra aquellos aspirantes que realizaron exactamente tres (3) intentos para ingresar a la UNAH, tal y como se muestra en la Figura 20 se aprecia claramente que la mayor motivación que tienen para utilizar sus intentos es debido a que no logran obtener los 700 puntos que solicita la UNAH sobre la PAA para ser admitidos y a medida realicen más intentos su probabilidad de aprobar es cada vez más baja.

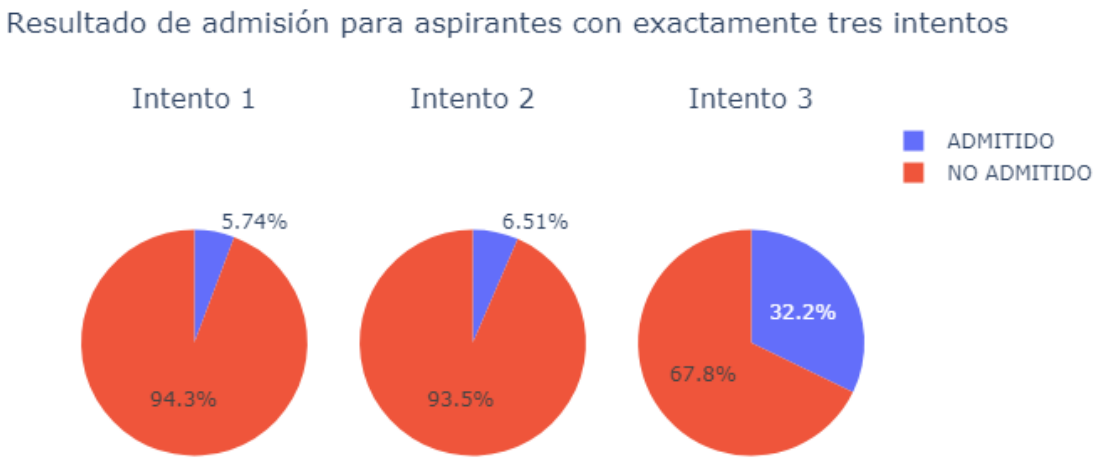

Figura 20 - Resultado de admisión para aspirantes con exactamente tres intentos (Fuente: elaboración propia)

En la Figura 21 se observa que, en la región sur, son más los aspirantes del sexo femenino que se someten a la PAA con la intención de estudiar una carrera universitaria. Además, del total de la muestra, son quienes obtienen mayormente un **no admitido** como resultado de realizar el examen de admisión.

Independientemente del origen del sector del instituto de estudio de educación media, los aspirantes tienden a tener mayor porcentaje de aprobación que reprobación en la Prueba de Aptitud Académica, tal como se muestra en la Figura 22.

Además, se observa que el mayor número de aspirantes son provenientes de institutos de educación media con carácter público, esto es debido a que el 94.86% de los estos son de familias numerosas de hasta siete (7) hermanos y con ingresos económicos familiares inferiores al rango de (L.) 15,000.01 a (L.) 20,000.00, donde el aspirante no aporta económicamente

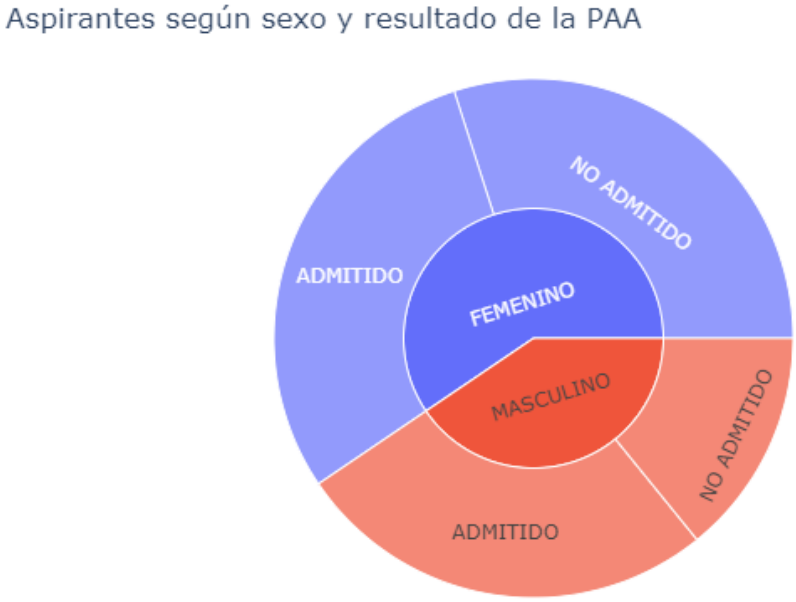

Figura 21 - Aspirantes según sexo y resultado de la PAA (Fuente: elaboración propia

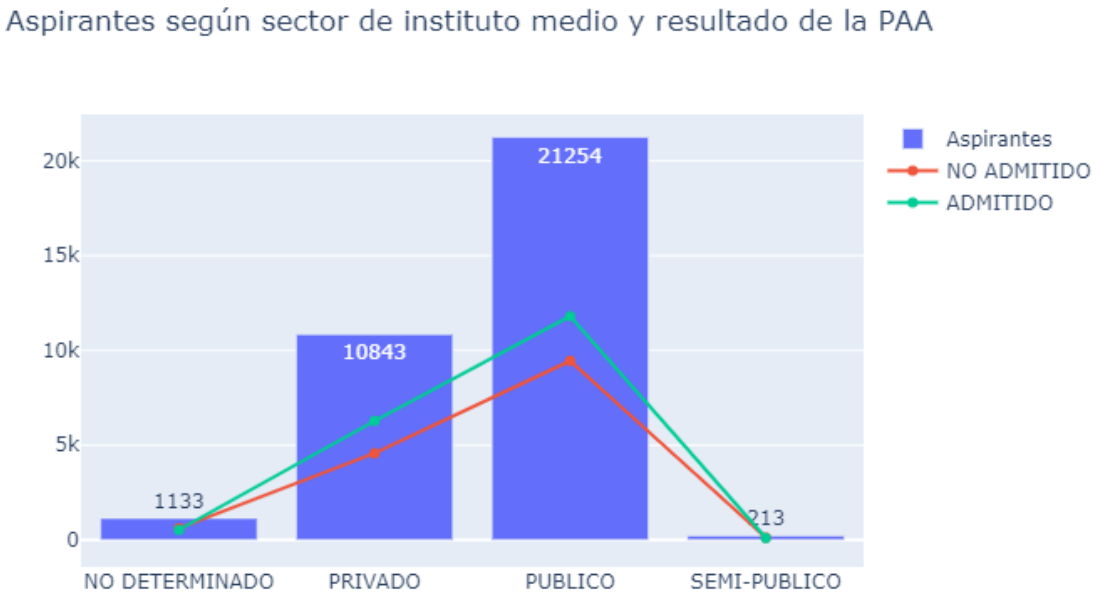

Figura 22 - Aspirantes según sector de instituto y resultado de la PAA (Fuente: elaboración propia)

En la Figura 23 se muestra como se divide el resultado de admisión según el puntaje obtenido en la PAA. La explicación de la Tabla 2 en la sección 2.1.4, indica que para ser admitido en la UNAH se necesita como requisito mínimo obtener 700 puntos en la PAA. Sin embargo, se determinan ciertos valores atípicos (*outliers*) con un resultado de **admitido** cuando debería ser **no admitido**.

Además, la Figura 24 muestra la relación del puntaje de la PAA verbal y matemática, considerando que la suma de ambas debe ser igual al puntaje obtenido, también se identificaron registros cuyo resultado de admisión no coincide con la premisa del puntaje mínimo de la PAA para ser o **no admitido** en la UNAH, y que la suma ambas difiere del puntaje de PAA obtenido.

En la sección 4.3 se deberá tratar estos valores ya sea indicando el resultado correcto que debería tener el aspirante o descartando del conjunto de datos de muestra aquellos registros que se consideren incorregibles y que generen ruido al momento de realizar el entrenamiento de los diferentes modelos de aprendizaje.

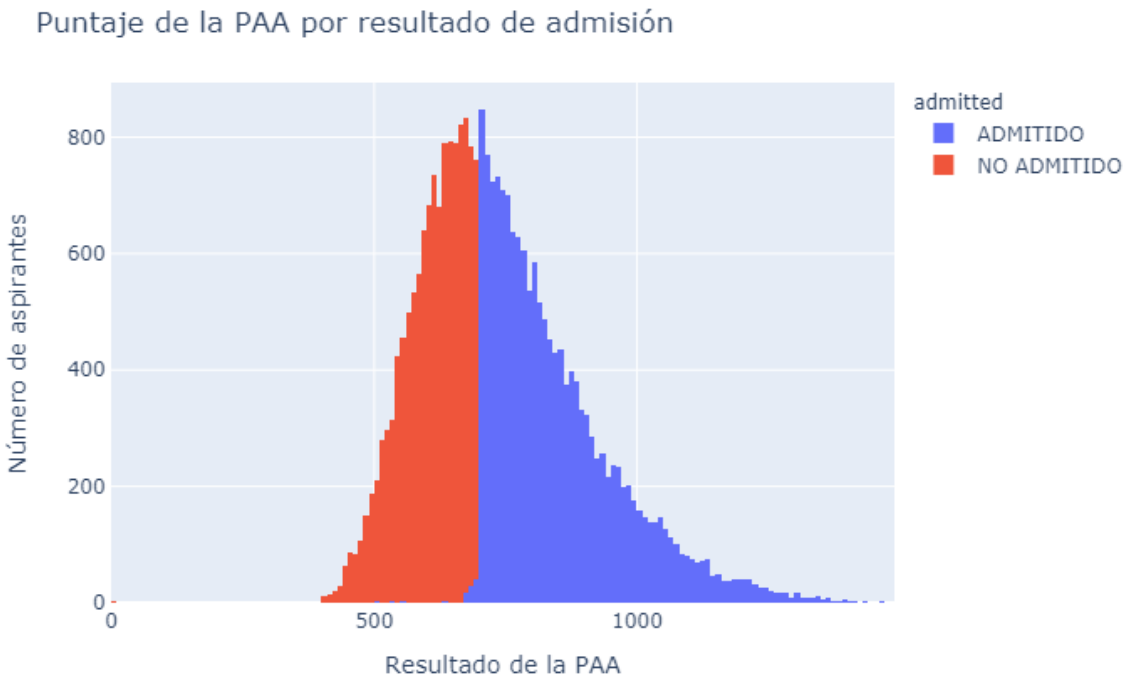

Figura 23 - Puntaje de la PAA por resultado de admisión (Fuente: Elaboración propia)

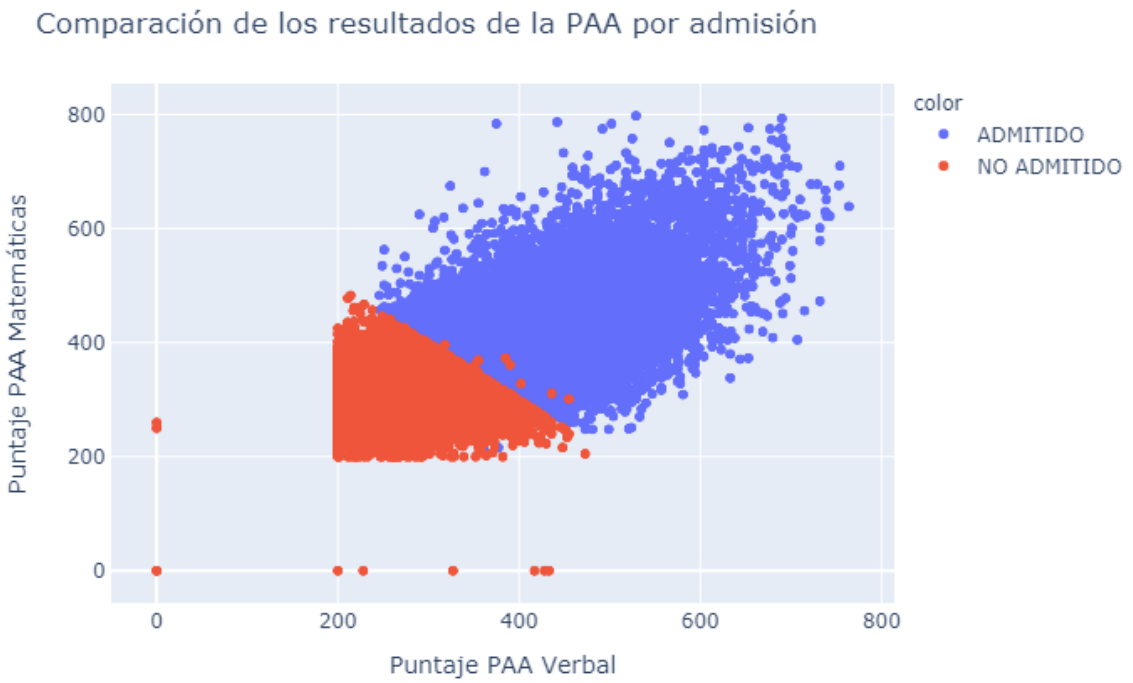

Figura 24 - Comparación de los resultados de la PAA por admisión (Fuente: elaboración propia)

4.3. Preparación de los datos

Se procede a construir el conjunto de datos final, el cuál será utilizado por los algoritmos o herramientas de modelado. En esta fase primeramente se realizará sobre el conjunto de datos la construcción de nuevas características y posteriormente la limpieza de los registros y campos que presenten anomalías o no sean significativos para el estudio.

Selección de datos

Esta fase de la metodología se realizará sobre el conjunto de muestra y validación previamente obtenido de forma aleatoria al comienzo de la comprensión de los datos, donde cada transformación y limpieza que se aplicará sobre ambos, excepto, aquellas acciones relacionadas con los puntajes obtenidos o resultados de admisión, donde únicamente se aplicará sobre el conjunto de muestra para asegurar la mejor calidad de los datos, pero no sobre el conjunto de validación para evitar el sobre ajuste de los modelos de aprendizaje.

Construcción de nuevos datos

El conjunto de datos originalmente proporcionado no tiene toda la información relevante para el objetivo planteado, por ello es importante identificar que campos o registros

se pueden generar a partir de la información existente. Además, que aporten valor al modelo de aprendizaje a generar. En el Anexo A – Ingeniería de características sobre el conjunto de datos, se muestran las acciones a realizar sobre el conjunto de datos de muestra con el objetivo de crear, transformar o reemplazar los campos que se poseen actualmente, todo cambio sobre este conjunto de datos se realizará también en el de validación.

En Figura 25 se muestra como la moda de edad de los aspirantes que realizan la PAA es de 19 años. Además, se ha determinado que existen registros cuya edad del aspirante es inferior a los 15 años, esto dependiendo de la frecuencia de los datos se puede considerar como un valor atípico en el conjunto de datos. Los aspirantes que realizan la PAA suelen ser estudiantes de educación media cuyo año de graduación coincide con la fecha de aplicación. Conociendo únicamente el año de graduación no se puede aglomerar los aspirantes a través de los años, el extremo derecho de la Figura 26 muestra que realmente realizan este proceso poco menos de un año después de su graduación. También se identifican ciertos registros cuyo lapso entre año de graduación y año de realización de la PAA es inferior a cero (0), todos aquellos inferiores a -1 años se consideran anómalos, dado que por premisa una vez aprobada la PAA el aspirante tiene únicamente un año para realizar la matrícula.

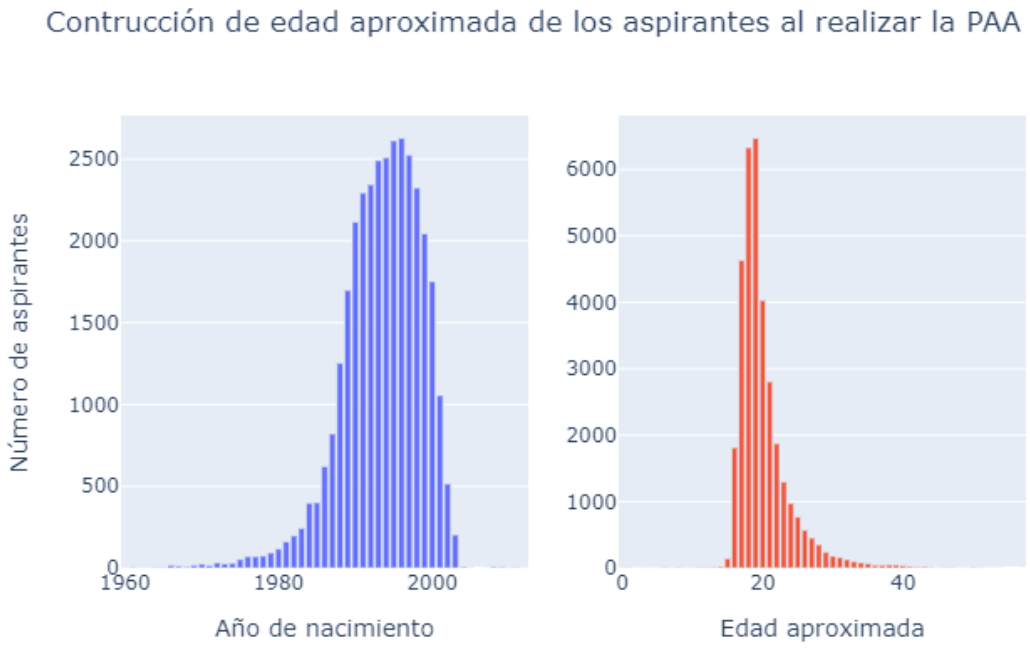

Figura 25 - Edad aproximada de los aspirantes al realizar la PAA (Fuente: elaboración propia)

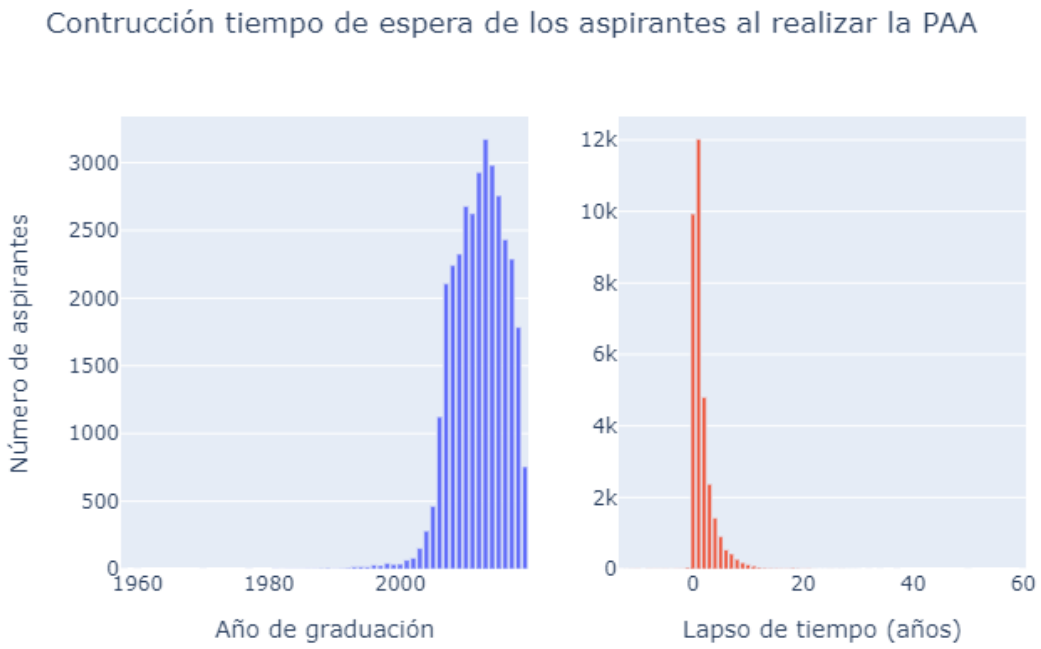

Figura 26 - Tiempo de espera de los aspirantes al realizar la PAA (Fuente: elaboración propia)

Dado que la PAA se realiza tres (3) veces al año como se mencionó anteriormente, se puede identificar mediante esta premisa los diferentes procesos que ha tenido la PAA por año, tal y como se muestra en la Figura 27. Así, se observa como la mayoría de los aspirantes han ido gradualmente a través de los años realizando la PAA en el primer y segundo proceso y como el tercer proceso poco a poco ha ido en decaimiento. Esto tiene sentido, ya que como se explicó en la Figura 26 la mayoría de los aspirantes suelen realizar el examen de admisión menos de un año posterior a la fecha de graduación.

El aspirante tiene la posibilidad de elegir tres (3) opciones de estudios de cualquier área. Sin embargo, ser **admitido** en la UNAH no implica ser admitido en una carrera, dado que cada carrera cuenta con sus propios criterios de admisión. En la Figura 28 se puede observar el resultado de la PAA por opción de estudios. Además, identificar la proporción de aspirantes que fueron admitidos en la primera, segunda y tercera opción seleccionada.

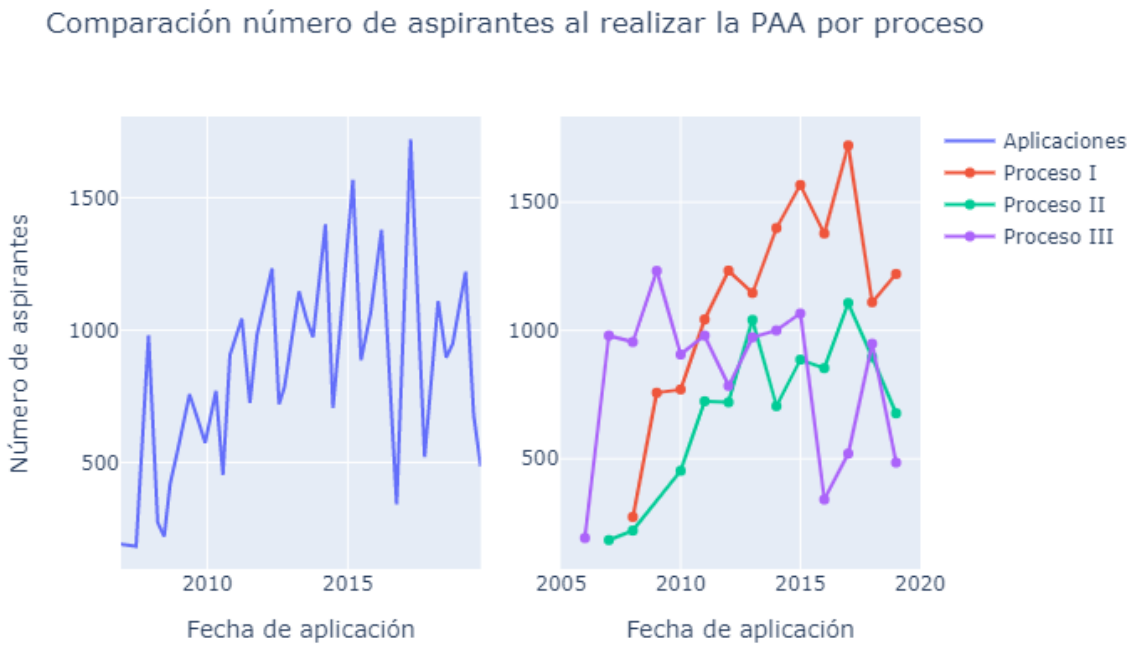

Figura 27 - Comparación número de aspirantes al realizar la PAA por proceso (Fuente: elaboración propia)

Figura 28 - Admisión de los aspirantes por opción de carrera de estudios (Fuente: elaboración propia)

Del total general, el 30.7% de los aspirantes son admitidos en su primera opción, el 11.9% en la segunda opción y el 11% en la tercera opción. La figura de la tercera opción muestra como existe un 2.32% de aspirantes que fueron **admitidos** a la UNAH y, sin embargo, no fueron admitidos en ninguna de las tres opciones anteriores, esto se atribuye probablemente a que no cuentan con el puntaje mínimo requerido para optar a alguna de las carreras. Este análisis tiene correspondencia con el realizado en la Figura 18.

Limpieza

Para mejorar el nivel de calidad de los datos a utilizar en modelados de aprendizaje automático (*machine learning*), es indispensable realizar una revisión y limpieza de valores atípicos dentro del conjunto de datos. Todas estas decisiones deben ser documentadas dado que se deben aplicar tanto para el conjunto de datos de muestra, como el de validación.

En primer lugar, se realizará una limpieza de los datos en base a los registros, posteriormente en base a los campos y por último se descarta la presencia de valores ausentes que quedaran en la construcción de nuevas características.

Cuando se intenta determinar la posición geográfica del lugar de nacimiento, residencia y dirección del instituto de los aspirantes que realizan la PAA, es probable que se generen valores nulos por encontrar '**no aplica (extranjero)**' en el campo de los departamentos.

En la Tabla 11 se muestran el porcentaje de registros que se convertirán en valores ausentes dentro de conjunto de datos y que por tanto se procederá a eliminarán del mismo para mantener la calidad de los datos.

Limpieza de registros

Tabla 11 – Porcentaje de registros descartados como limpieza en campos de dirección

Campo	Valor atípico	Porcentaje (%)
birth_department	no aplica (extranjero)	0.09
residency_department	no aplica (extranjero)	0.05
institute_department	no aplica (extranjero)	0.09

En la Figura 29 se muestra el resultado de realizar la limpieza de los registros identificando si, el resultado de admisión coincide con el puntaje mínimo requerido para aprobar la Prueba de Aptitud Académica.

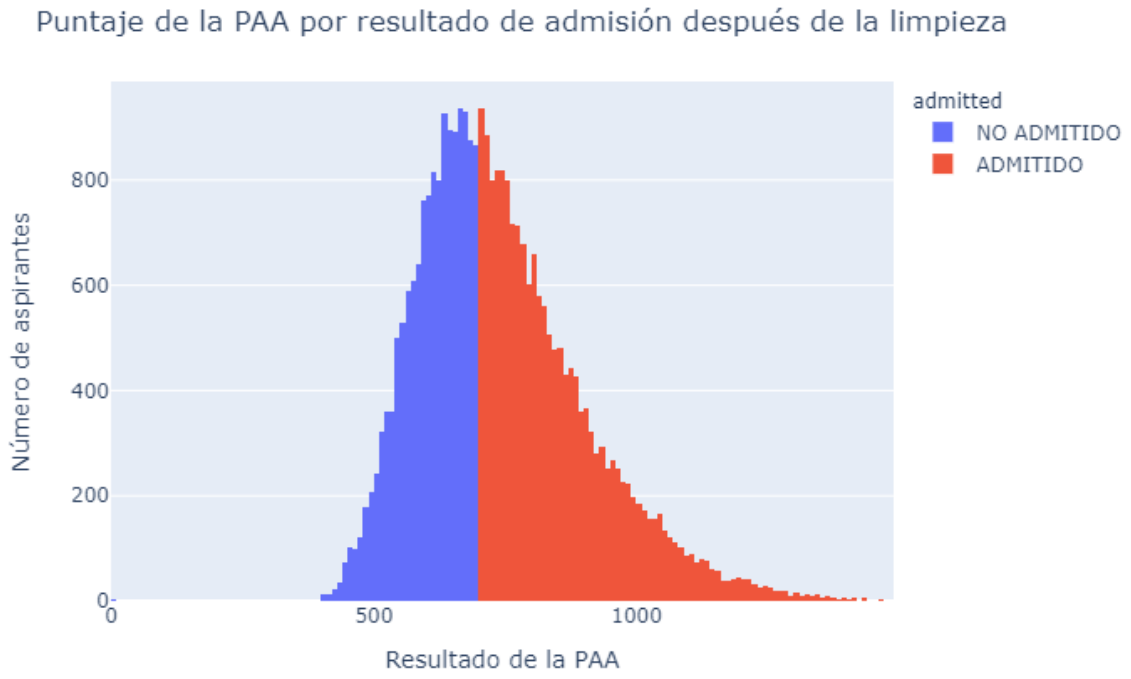

Figura 29 - Puntaje de la PAA por resultado de admisión después de la limpieza (Fuente: elaboración propia)

Observando la Figura 23 se determinó que existían registros dentro del conjunto de datos proporcionados donde, la suma de la parte verbal y matemática no coincide con el puntaje obtenido de la PAA, cuando por premisa, la suma de ambas partes debe coincidir con el puntaje obtenido, por ello se procedió a realizar una limpieza a los datos descartando aquellos que no cumplen esta premisa.

La Figura 30 muestra la limpieza como resultado de esta comparación, en donde, a diferencia de la Figura 24 no se presentan errores de etiquetado en el diagrama de dispersión. También se aprecia un corte notable a los 200 puntos tanto para el eje en la parte verbal, como el eje en la parte matemática, esto sucede porque como se comentó en la sección 2.1.4 este es el valor mínimo por cada parte (verbal y matemática) que puede adquirir un aspirante que realiza la PAA.

Además, se aprecian registros que están por debajo de estos valores, esto es porque algunos aspirantes solo realizan la PCCNS y PAM como requisito para ser admitidos en

alguna carrera del área médica o ingeniería respectivamente. En general, después de la limpieza los resultados de admisión, tanto al comparar el puntaje de la PAA obtenido, así como la suma de la parte verbal y matemática, no se entrelazan generando valores atípicos.

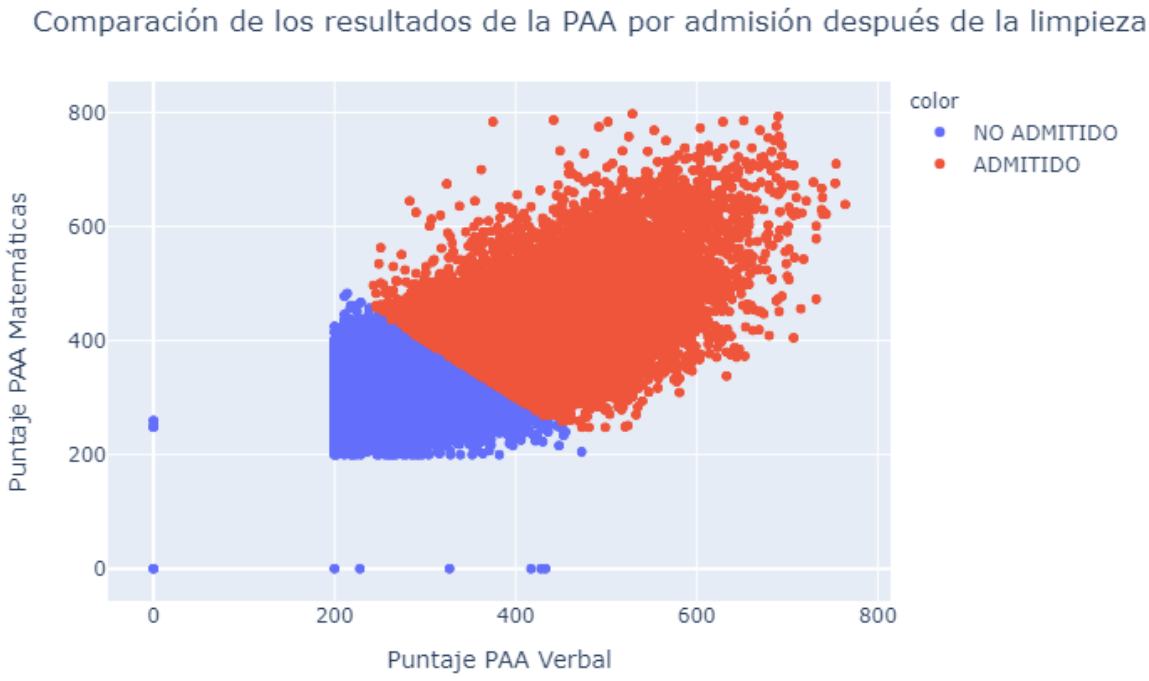

Figura 30 - Comparación de los resultados de la PAA por admisión después de la limpieza (Fuente: elaboración propia)

En el extremo izquierdo de la Figura 31 se aprecia como al realizar el cálculo de la edad aproximada del aspirante el diagrama de cajas muestra, en el extremo superior de la del bigote de la caja, la presencia de valores atípicos. Contar con datos limpios y correctamente etiquetados es indispensable para obtener buenos resultados de los modelos. Para evitar el descarte completo de estos registros se reemplazan dichos valores por la mediana de la edad en el conjunto de datos, ya que esta medida es menos sensible a oscilaciones de los valores de la variable en comparación con la media cuya expresión es afectada por los valores extremos. Además, la mediana es más representativa que la media en este caso, ya que se está considerando una variable bastante heterogénea.

Originalmente, la edad aproximada en los datos proporcionados tiene un rango que va de 0 hasta 57 años, claramente es imposible que un aspirante tenga cero (0) años al momento de realizar la PAA, esto se da porque el campo que permite la construcción de la edad aproximada (**year_birth**) se rellena manualmente. Sin embargo, nada evita que una

persona de 57 años, mientras cumpla haber terminado los estudios de educación media, se someta a la PAA. Aun así, se corrigen los valores extremadamente atípicos acotando las edades en un intervalo que va de 10.5 y 28.5 años, y reemplazando aquellos que no estén en este intervalo por 19 años que es el valor de la mediana para este campo.

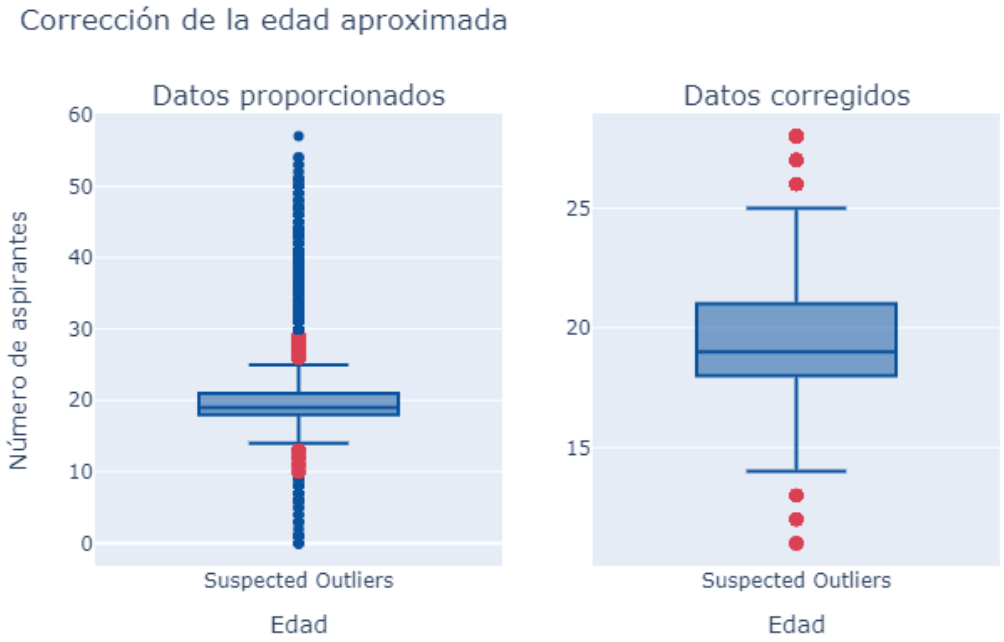

Figura 31 - Corrección de la edad aproximada (Fuente: elaboración propia)

Igual que en el caso anterior, en la Figura 32 se procedió a realizar una corrección del lapso en año entre el examen de admisión y graduación del aspirante. Se determinan lapsos negativos, esto es posible dado que, por premisa, el aspirante una vez siendo **admitido** en la UNAH tienen un año para tramitar la matrícula, ese año puede ser el que le resta para finalizar los estudios de educación media. El proceso de admisión descrito en la sección 2.1.3, explica que la documentación solicitada para realizar la PAA es únicamente: un documento identificativo y el recibo de pago realizado. Para la matrícula, si es indispensable contar con un documento acreditativo de haber finalizado los estudios de educación media.

El lapso es determinado utilizando el campo del año de graduación (**graduation_year)** cuyo resultado tiene un rango que va -13 a 34 años, igual que en el caso de la edad aproximada, nada impide que una persona espere varios años luego de haber sido graduado de educación media para someterse a la PAA. Aun así, se corrigen los valores

extremadamente atípicos determinados a través de la Ecuación 2 del diagrama de cajas y bigotes, acotando el lapso en un intervalo que va de -3 y 7, y reemplazando aquellos registros que no estén en este intervalo por 1 año, dado que este es el valor de la mediana para este campo, el cuál es representativo considerando la heterogeneidad del dato.

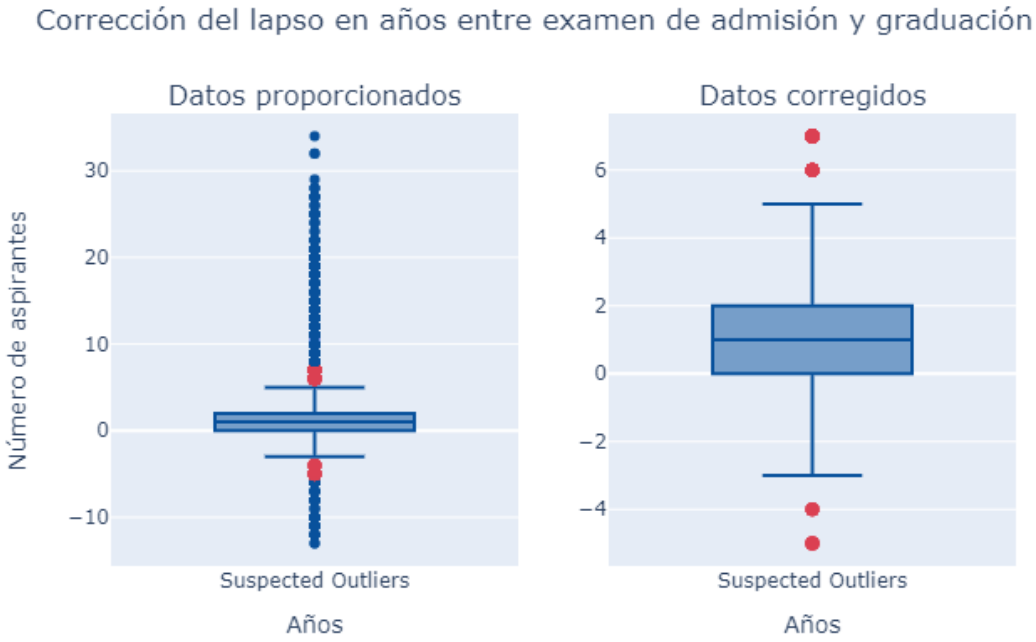

Figura 32 - Corrección del lapso en años entre examen de admisión y graduación (Fuente: elaboración propia)

También en la Figura 33 se utiliza el diagrama de cajas para describir la distribución del número de hermanos que tiene el aspirante al momento de realizar el examen de admisión. Se observan aspirantes con familias muy numerosas. Sin embargo, al aplicar la Ecuación 2 explicada en la sección 2.2, se determinan valores extremadamente atípicos cuando el número de hermanos es superior a 9.

En los datos proporcionados el número de hermanos se encuentra en un intervalo que van de 0 a 15, por ello se corrigen estos valores acotando dicho número de 0 a 9 y reemplazando el resto de los registros por 3, dado que este es el resultado de la mediana.

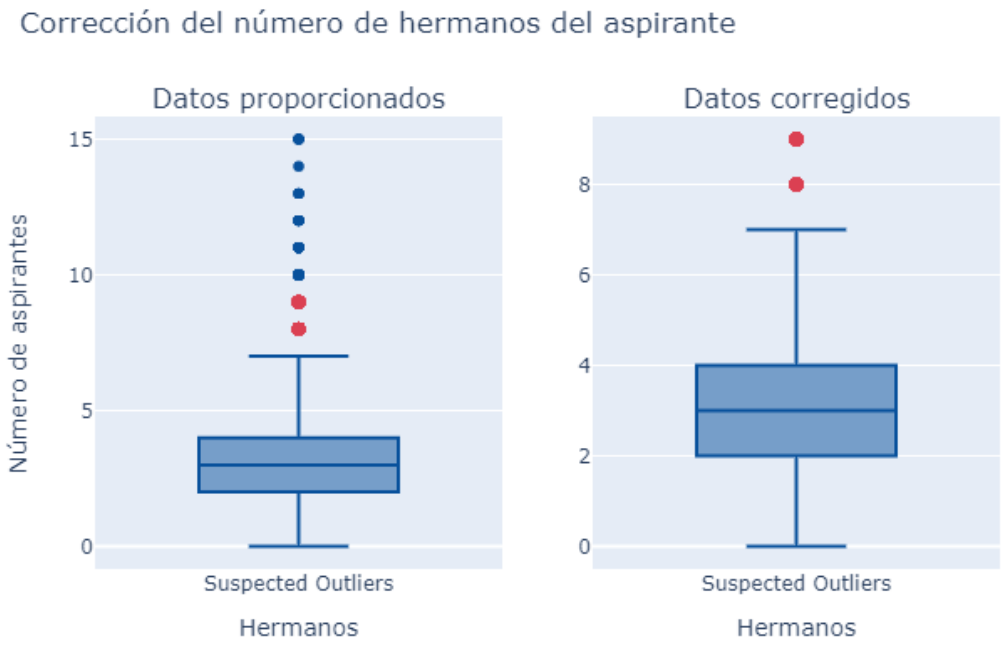

Figura 33 - Corrección del número de hermanos del aspirante (Fuente: elaboración propia)

Limpieza de campos

Con anterioridad en esta misma etapa de la metodología, se realizó la creación de campos nuevos y transformación de los ya presentes, se procede a seleccionar únicamente aquellos campos de carácter numérico y la variable objetivo para proceder al modelado.

Se ha considera utilizar únicamente los datos numéricos, ya que la dificultad que presentan los campos categóricos consiste en el número de posibles valores únicos que estos posean. Al momento de generar el modelo, se deberá distinguir los campos característicos o independientes y el campo objetivo o dependiente. Los campos del tipo categóricos deberán convertir cada posible valor único en columnas o realizar una indexación (convertir el dato no numérico en numérico). Considerando que el modelo se está generando con datos de la región sur del país, algunos campos se convertirán en conflictivos (p. ej, el nombre del instituto **institute_name,** donde el modelo que se genere actualmente contendría información que no es común a las demás regiones del país, por lo que no podría ser generalizado a toda la organización). Además, aumentaría la complejidad de procesamiento por lo que se considera inviable.

El conjunto de datos final previo al modelado consiste en las siguientes características que contiene la Tabla 12, se aprecian que el número de campos se ha reducido a 40 lo que representa un 74.04% del original, en cuanto al número de registros, el conjunto de validación se sometió al mismo proceso de limpieza que la muestra excepto en la limpieza del admisión_rate erróneas, esto debido a que los datos de validación deben desconocer tratamientos referidos al resultado de realizar la PAA, si no fuese así se estaría realizando un sobreajuste en el modelo final.

Tabla 12 - Características del conjunto de datos final

	Observación
Número de campos	• 39 numéricos • 1 objetivo (categórico)
Número de registros	• Muestra: 26,527 (93.31%) • Validación: 4,851 (96.69%)

4.4. Modelado

Esta fase implica la selección del modelado a generar, así como la configuración de los parámetros para conseguir resultados óptimos, lo que conlleva a realizar ajustes adicionales de "preparación de datos" dentro del conjunto de datos final.

Técnicas de modelado

Como se comentó en la comprensión del negocio, es importante identificar los factores que juegan un papel fundamental para que un aspirante apruebe o no el examen de admisión y pueda ser **admitido** en la UNAH. El campo objetivo es del tipo de dato categórico, por lo que se utilizarán técnicas de aprendizaje automático supervisado, específicamente de clasificación para su predicción. Se realiza una comparativa entre los algoritmos **Decisión Tree Classifier** (perteneciente a la clase ***tree*** de la librería ***sklearn)***, Random Forest Classifier (perteneciente a la clase ***ensemble*** de la librería ***sklearn***) y XGBoost Classifier (perteneciente a la librería ***xgboost***) para determinar cual genera los mejores resultados y por tanto ser utilizado como modelo final que sirva de apoyo a las estrategias que actualmente elaboran en la Dirección del Sistema de Admisión (DSA).

Para la construcción de las matrices de confusión, como métrica de evaluación en los modelos de clasificación, se identificarán como valor verdadero positivo los aspirantes cuyo resultado de admisión es **admitido** y como valor verdadero negativo los aspirantes que hayan obtenido un **no admitido**.

Diseño de las pruebas

Antes de iniciar la comprensión de los datos, se dividió el conjunto de datos en muestra y validación con un 85% y 15% respectivamente. Para determinar el modelo final se utilizarán estos conjuntos de datos, pero ya efectuada la limpieza mencionada.

Se realizará una validación cruzada utilizando ***GridSearchCV*** de la clase ***model_selection*** perteneciente a ***sklearn*** que varíe los parámetros de configuración en los algoritmos de aprendizaje automático y seleccionando el mejor modelo utilizando la precisión (accuracy) como métrica de evaluación. Al generar los modelos, la muestra se dividirá de manera aleatoria, mediante la librería ***train_test_split*** perteneciente a ***sklearn***, de igual manera en un 85% submuestra (*train*) y 15% prueba (*test*) tal y como se muestra en la Tabla 13.

Tabla 13 - Esquema de la división del conjunto de muestra en train/ test para la validación cruzada

	Registros	Campos
Conjunto de muestra	26527	40
Train	22547	40
Test	3980	40

Una vez identificado el modelo con la mejor métrica de evaluación para cada algoritmo, se aplicará sobre este modelo, el conjunto de validación para describir las capacidades de este en base a los resultados que se obtengan. Además, se dispondrá de información sobre el recall y f1_score para cada modelo. La validación cruzada permitirá identificar la mejor configuración de parámetros para el modelo Random Forest Classifier, Decision Tree Classifier, XGBoost Classifier, tal y como se muestra en las Tabla 14 y Tabla 15:

Tabla 14 - Parámetros de evaluación en validación cruzada para los modelos

Modelos de aprendizaje automático	Parámetros
Random Forest Classifier	n_estimators, max_depth
Decisión Tree Classifier	criterion, max_depth
XGBoost Classifier	n_estimators, criterion, max_depth

Tabla 15 - Variación de los parámetros de los modelos

Parámetro	Variación
• **n_estimators**: Número de árboles en el bosque.	**Rango: [70, 95]** **Intervalo: 5**
• **criterion**: Función para medir la calidad de las divisiones. o **gini**: tenderá a encontrar la clase más grande. o **entropy**: tiende a encontrar grupos de clases que constituyen ~ 50% de los datos.	**Rango: [gini, entropy]**
• **max_depth**: La profundidad máxima del árbol. La profundidad del árbol implica la utilización de más nodos de decisión (campos o características) en el árbol binario	**Rango: [5, 30]** **Intervalo: 5**

La variación planteada para cada parámetro en la Tabla 15 fue determinada bajo experimentación, donde se realizó la ejecución de los diferentes modelos y se encontraron los rangos e intervalos que generaron los mejores resultados en los diferentes estos, no se muestran todas las pruebas para que el documento sea más legible.

Construcción del modelo

Ahora, se procederá a ejecutar los algoritmos de modelado sobre el conjunto de datos de muestra, recordando que éste se dividirá en conjuntos de entrenamiento y prueba:

Decision Tree Classifier

En la Tabla 16 se muestran las diferentes configuraciones utilizadas para determinar el mejor modelo en la validación cruzada con la técnica Decision Tree Classifier de la clase *tree* perteneciente a la librería *sklearn*. Se especificaron dos (2) tipos de criterios y cinco (5) diferentes profundidades máximas del árbol, así se han obtenido 10 modelos de predicción.

La configuración óptima se determinó en el primer modelado con un criterion: gini y un max_depth: 5 generando la mejor precisión de 63.21%.

Tabla 16 - Validación cruzada Decision Tree Classifier

	criterion	max_depth	mean_test_score	rank_test_score
1	gini	5	63.21	1
2	gini	10	62.50	4
3	gini	15	60.60	6
4	gini	20	58.41	8
5	gini	25	58.04	10
6	entropy	5	63.07	2
7	entropy	10	63.04	3
8	entropy	15	60.61	5
9	entropy	20	59.43	7
10	entropy	25	58.36	9

En la Figura 34 se muestra la comparación del *criterion* y *max_depth* para los 10 modelos en la validación cruzada utilizando Decision Tree Classifier. Se aprecia como a medida se aumenta la profundidad del árbol, también aumenta el ruido que generan las nuevas variables características adicionadas, tal que la precisión comienza a disminuir drásticamente. Cabe mencionar que el rango de la precisión en la Figura 34 va de [0.58, 0.63] ya que si fuese de [0, 100] no se podría apreciar la comparativa.

La Figura 35 muestra en orden los campos o características que explican en un 95% el modelo Decisión Tree Classifier final. En este caso se aprecia que solo se necesitan 5 campos para determinar si un aspirante será o no **admitido** en la UNAH; el campo más importante para asegurar esto, es conocer el puntaje mínimo requerido para ser admitido en la carrera de la primera opción.

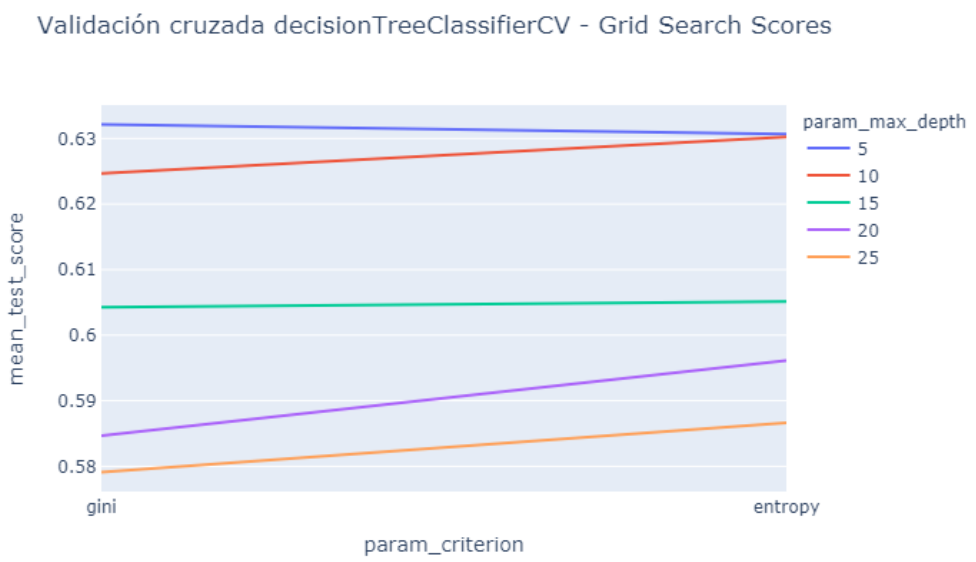

Figura 34 - Validación cruzada Decision Tree Classifier (Fuente: elaboración propia)

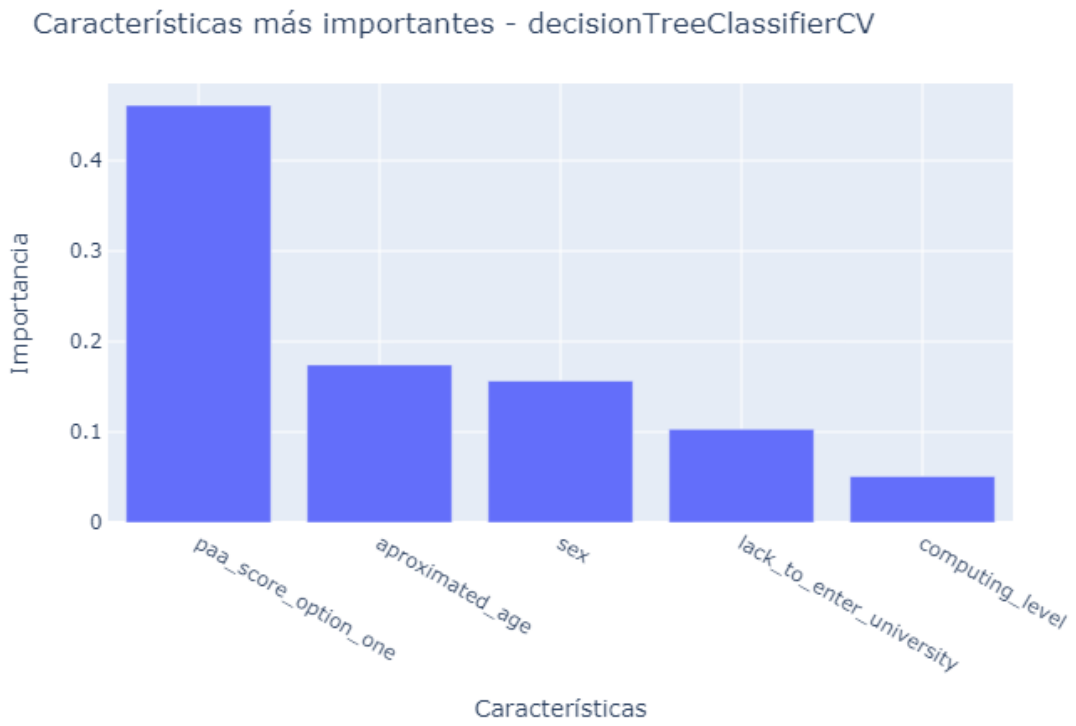

Figura 35 - Características más importantes, Decision Tree Classifier (Fuente: elaboración propia)

Ahora, que se ha determinado la mejor configuración para el Decision Tree Classifier, se describirá el rendimiento del modelo supervisado en los datos de prueba. En la Tabla 17 se ve como la proporción que clasifica como verdadero positivo y negativo es mayor que los falsos positivos y negativos, esta es una precisión moderadamente alta considerando un 100% de precisión como resultado esperado.

Tabla 17 - Matriz de confusión, datos de prueba Decision Tree Classifier

		Predicción		**Total, realidad**
		ADMITIDO	NO ADMITIDO	
Realidad	ADMITIDO	1480	703	2183
	NO ADMITIDO	782	1015	1797
Total, predicción		2262	1718	

El reporte de la clasificación en la Tabla 18 muestra como existe una sensibilidad (recall) aceptable para determinar los aspirantes que serán **admitidos** en la UNAH, mientras que su especificidad no llega ni el 60%. Considerando que el estudio consiste en identificar los aspirantes que **no lograrán ser admitidos**, será necesario que el modelo sea capaz de predecir con mejor precisión los verdaderos negativos.

Tabla 18 - Reporte de la clasificación, Decision Tree Classifier

	precision	**recall**	**f1-score**	**support**
ADMITIDO	0.65	0.68	0.67	2183
NO ADMITIDO	0.59	0.56	0.58	1797
accuracy			0.63	3980
macro avg	0.62	0.62	0.62	3980
weighted avg	0.63	0.63	0.63	3980

Random Forest Classifier

Se realizaron 50 modelos en la validación cruzada para determinar la mejor configuración utilizando Random Forest Classifier de la clase *ensemble* perteneciente a la librería *sklearn*. En esta ocasión, en el Anexo D – Validación cruzada Random Forest

Classifier, se muestra como se especificaron dos (2) tipos de criterios de subdivisión del conjunto de entrenamiento, cinco (5) diferentes profundidades máximas de los árboles del bosque y cinco (5) diferentes estimadores o número de árboles en el bosque de decisión, con lo cual se generaron 50 modelos de predicción.

La configuración óptima se determinó en el modelo número 38 con *criterion*: *entropy*, *max_depth*: 15 y *n_estimators*: 80 generando la mejor precisión de 65.91%.

En la Figura 36 se compara la media de la precisión en base al número de estimadores y la profundidad máxima de los árboles del bosque de decisión. Previamente se filtraron los 50 modelos para el mejor criterio con el objetivo de hacer más legible la información en la figura. El rango de la precisión se expresa de [0.64, 0.66] para que se aprecie la comparativa de las diferentes configuraciones.

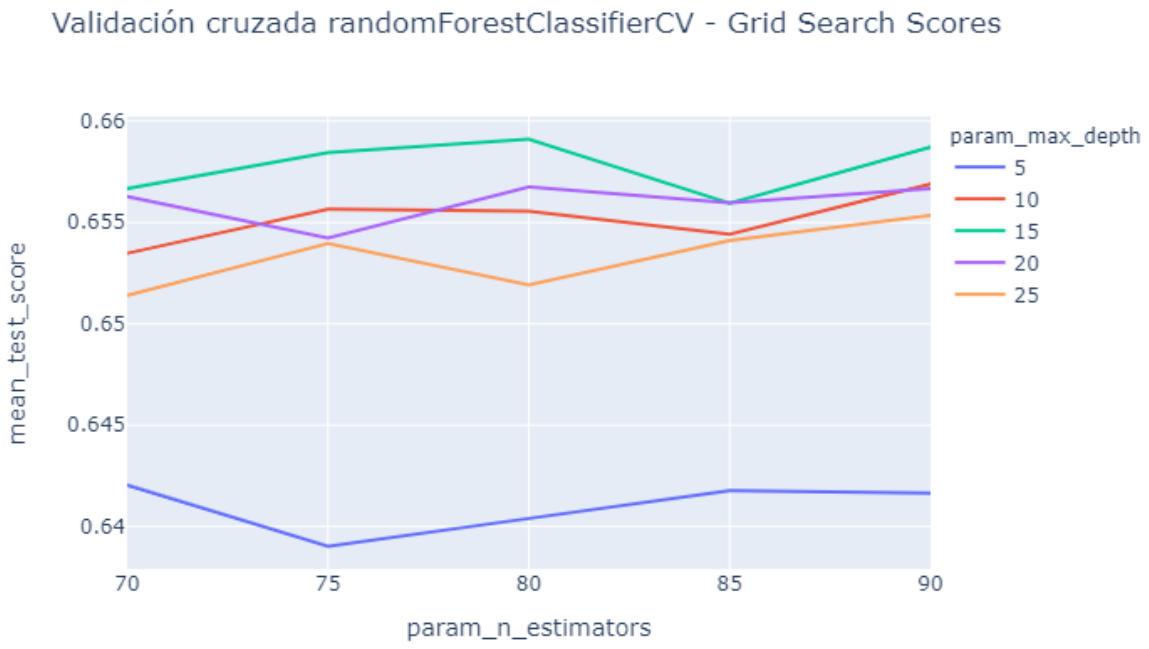

Figura 36 - Validación cruzada Random Forest Classifier (Fuente: elaboración propia)

En la Figura 37 se muestra de forma ordenada los campos o características que explican en un 95% el modelo Random Forest Classifier final. En esta ocasión, se necesitan muchos más campos para explicar el modelo, en otras palabras, a medida se van adicionan

campos o características al modelo aumenta la profundidad del árbol y el algoritmo maneja mejor el ruido que estos campos pueden generar a diferencia del Decision Tree Classifier. Sin embargo, los campos importantes del algoritmo anterior pertenecen a las primeras seis (6) características del Random Forest Classifier, pero en esta ocasión siendo la edad aproximada la más importante.

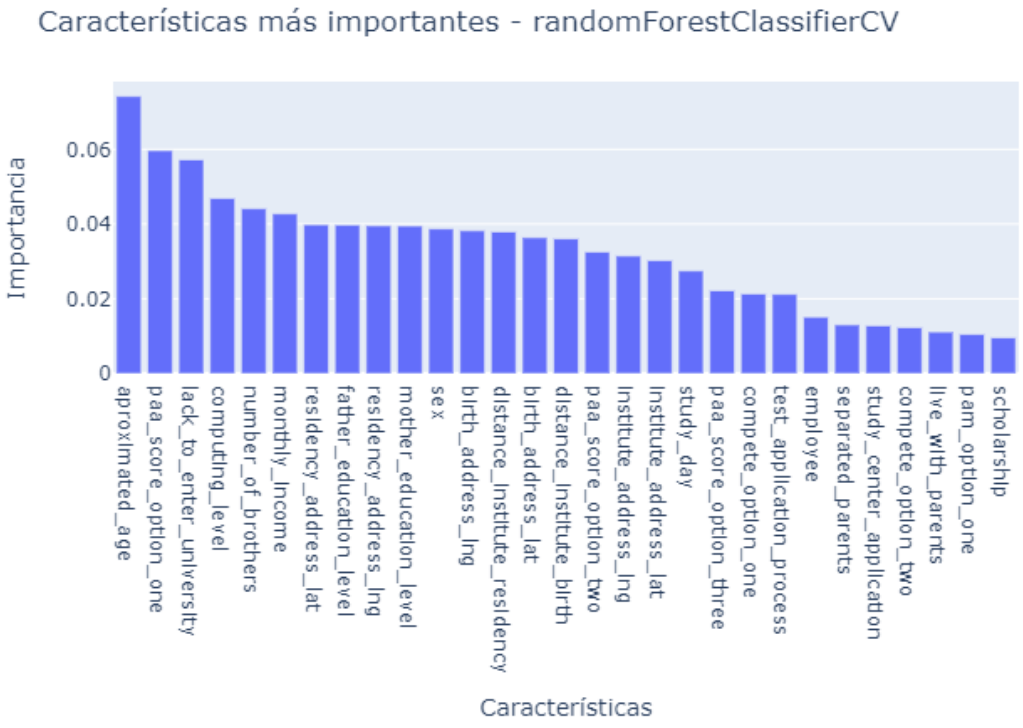

Figura 37 - Características más importantes, Random Forest Classifiert (Fuente: elaboración propia)

En la Tabla 19 se describe el rendimiento del modelo Random Forest Classifier final, obtenido después de la validación cruzada, utilizando los datos de prueba. En comparación al Decisión Tree Classifier, se ve como los falsos positivos y falsos negativos han disminuido considerablemente, por lo que se puede asegurar que este modelo es mejor que el generado por el algoritmo anterior.

Tabla 19 - Matriz de confusión, datos de prueba Random Forest Classifier

		Predicción		**Total, realidad**
		ADMITIDO	NO ADMITIDO	
Realidad	ADMITIDO	1512	671	2183
	NO ADMITIDO	722	1075	1797
Total, predicción		2234	1746	

En la Tabla 20 se aprecia como la sensibilidad para determinar si un aspirante será **admitido** ha mejorado en comparación al Decisión Tree Classifier. Sin embargo, su especificidad apenas alcanza el 60%.

Tabla 20 - Reporte de la clasificación, Random Forest Classifier

	precision	recall	f1-score	support
ADMITIDO	0.68	0.69	0.68	2183
NO ADMITIDO	0.62	0.60	0.61	1797
accuracy			0.65	3980
macro avg	0.65	0.65	0.65	3980
weighted avg	0.65	0.65	0.65	3980

XGBoost Classifiert

Como último modelo de comparación, en el Anexo E – Validación cruzada XGBoost Classifier se muestran las configuraciones utilizadas para determinar el mejor modelo utilizando XGBoost Classifier perteneciente a la librería *xgboost*. En esta ocasión, se especifican dos (2) tipos de criterios, cinco (5) diferentes niveles de profundidad máxima de los árboles y cinco (5) diferentes y cinco (5) diferentes estimadores o números de árboles en el bosque de decisión, con esto se generaron 50 modelos de predicción. Al realizar la validación cruzada se determinó como mejor modelo el de la iteración 2 con *criterion: gini, max_depth*: 5, *n_estimators*: 75 y una precisión del 66.07%.

Se compara gráficamente en la Figura 38 los diferentes modelos de la validación cruzada utilizando XGBoost. De los 50 modelos, se realizó un filtro extrayendo únicamente los del criterion gini, esto para hacer más legible la información en la gráfica. En esta ocasión el rango de la precisión se encuentra de [0.64, 0.66].

En la Figura 39 se observa de forma ordenada los campos o características que mejor explican en un 95% el modelo XGBoost Classifier final. Al igual que el Random Forest Classifier, necesita muchos más campos para explicar el modelo que el Decision Tree Classifier. Además, los tres (3) algoritmos coinciden en sus primeras características.

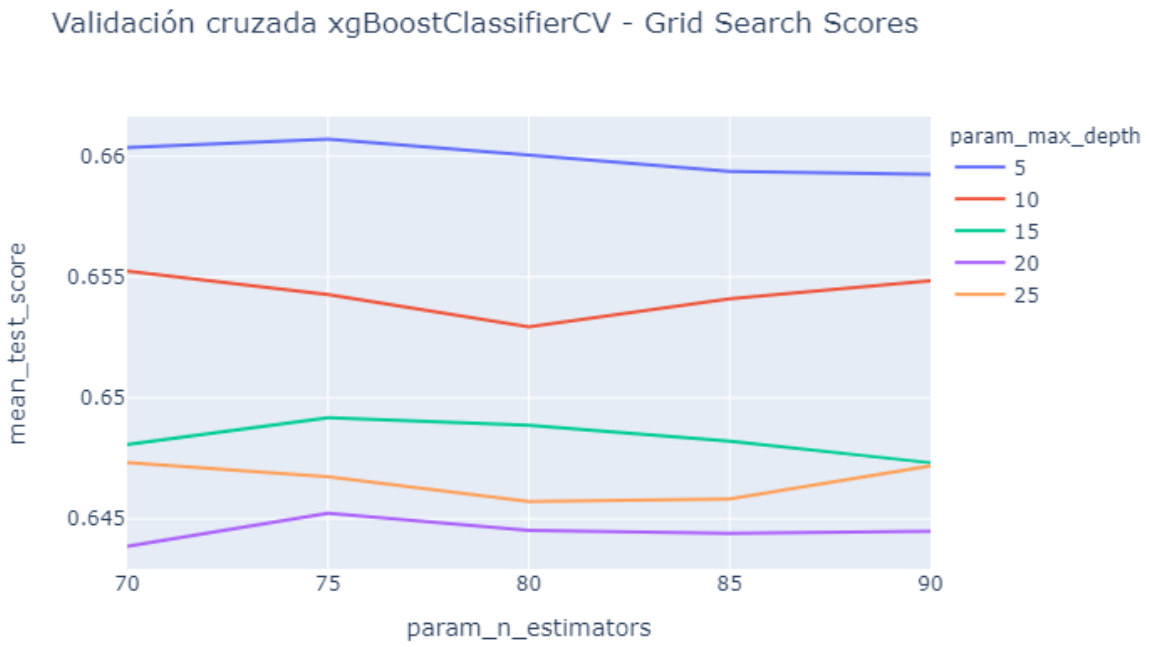

Figura 38 - Validación cruzada XGBoost Classifier (Fuente: elaboración propia)

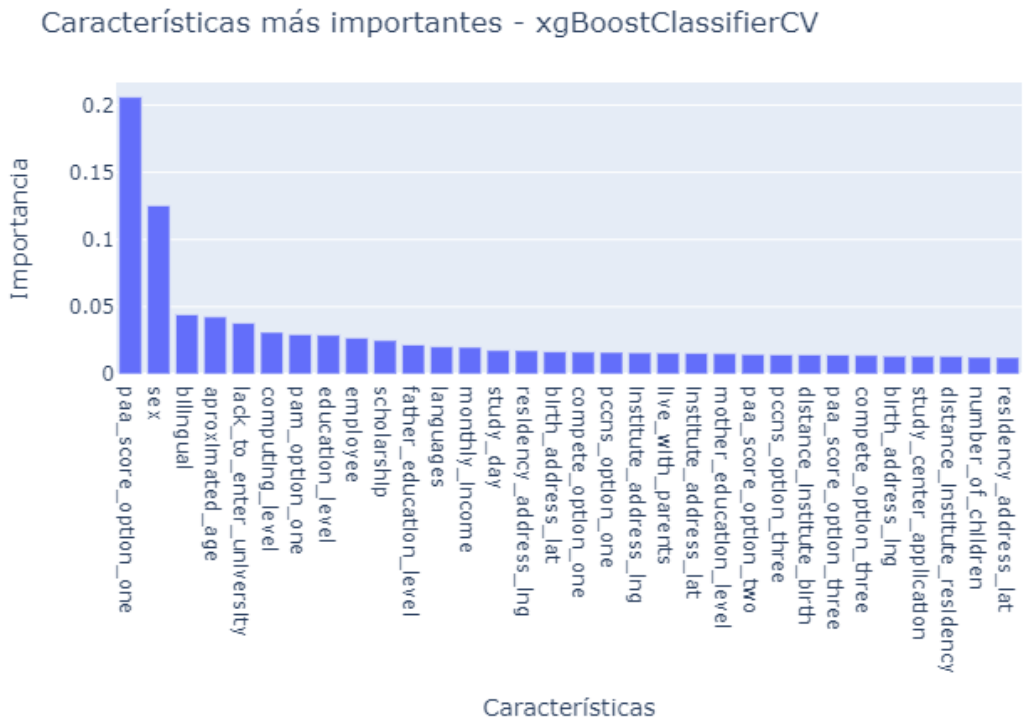

Figura 39 - Características más importantes, XGBoost Classifier (Fuente: elaboración propia)

Al evaluar el modelo XGBoost Classifier con los datos de prueba se observa en la Tabla 21 que el número de falsos positivos creció y que el número de falsos negativos disminuyó en comparación al Random Forest Classifier, utilizando los mismos datos de prueba para ambos modelos.

Tabla 21 - Matriz de confusión, datos de prueba XGBoost Classifier

		Predicción		**Total, realidad**
		ADMITIDO	NO ADMITIDO	
Realidad	ADMITIDO	1540	643	2183
	NO ADMITIDO	732	1065	1797
Total, predicción		2272	1708	

En la Tabla 22 se observa como la sensibilidad para determinar la admisión de un aspirante ha aumentado en comparación al Random Forest Classifier. Sin embargo, la especificidad se ha disminuido a un 59%. Aun así, al comparar el F1-Score en ambos modelos se ve como la precisión del recall ha aumentado en XGBoost, mientras que la de la especificidad ha mantenido, en otras palabras, la capacidad del modelo para identificar a los aspirantes vulnerables es la misma en comparación al Random Forest Classifier.

Tabla 22 - Reporte de clasificación, XGBoost Classifier

	precision	**recall**	**f1-score**	**support**
ADMITIDO	0.68	0.71	0.69	2183
NO ADMITIDO	0.62	0.59	0.61	1797
accuracy			0.65	3980
macro avg	0.65	0.65	0.65	3980
weighted avg	0.65	0.65	0.65	3980

En la Tabla 23 se muestra un resumen de la comparativa entre los tres modelos, determinando que, a pesar de su precisión, el modelo Random Forest Classifier con configuración criterion: entropy, max_depth: 15 y n_estimators 85, es el que será utilizado como modelo final para la detección temprana de aspirantes vulnerables que intentan

ingresar a la UNAH, ya que posee la misma capacidad de identificación de aspirantes vulnerables que le XGBoost y además, su tiempo de entrenamiento es más corto.

Tabla 23 - Comparación de los modelos de clasificación

Modelo	Mejor configuración			Precisión (Accuracy)		Tiempo de entrenamiento
	criterion	max_depth	n_estimators	Entrenamiento	Prueba	
Decision Tree Classifier	gini	5		63.21	62.69	0:00:06
Random Forest Classifier	entropy	15	80	65.91	65.38	0:05:20
XGBoost Classifier	gini	5	75	66.07	65.45	0:14:56

Evaluación técnica

Una vez determinado el mejor modelo según la métrica establecida, se procede a describir su rendimiento con el conjunto de datos de validación. La matriz de confusión en la Tabla 24 se describe el nivel de clasificación que posee el modelo final utilizando el conjunto de datos de validación o completamente desconocido.

Tabla 24 - Matriz de confusión, conjunto de datos de validación, Ramdon Forest Classifier

		Predicción		Total, realidad
		ADMITIDO	NO ADMITIDO	
Realidad	ADMITIDO	1872	843	2715
	NO ADMITIDO	817	1319	2136
Total, predicción		2689	2162	

En la Tabla 25 se observa que la sensibilidad del modelo para determinar aspirantes que serán **admitidos** aumenta un poco utilizando los datos de validación en comparación a los datos del conjunto de prueba en la selección del mejor modelo, mientras que la especificidad ha disminuido. Aunque, revisando el F1-Score, la precisión para la sensibilidad y especificidad se ha mantenido tanto en la generación y selección del modelo como en la validación utilizando el conjunto de datos separado al inicio.

Tabla 25 - Reporte de clasificación, conjunto de datos de validación, Ramdon Forest Classifier

	precision	recall	f1-score	support
ADMITIDO	0.69	0.69	0.69	2715
NO ADMITIDO	0.61	0.61	0.61	2136
accuracy			0.66	4851
macro avg	0.65	0.65	0.65	4851
weighted avg	0.66	0.66	0.66	4851

En la Figura 40 se muestra la curva ROC para la validación del Random Forest Classifier utilizando los datos del conjunto de validación. Se aprecia como 1 – specificy que determina el punto de corte más alto para la sensibilidad, es un valor muy cercano al en el reporte de la clasificación de la Tabla 25.

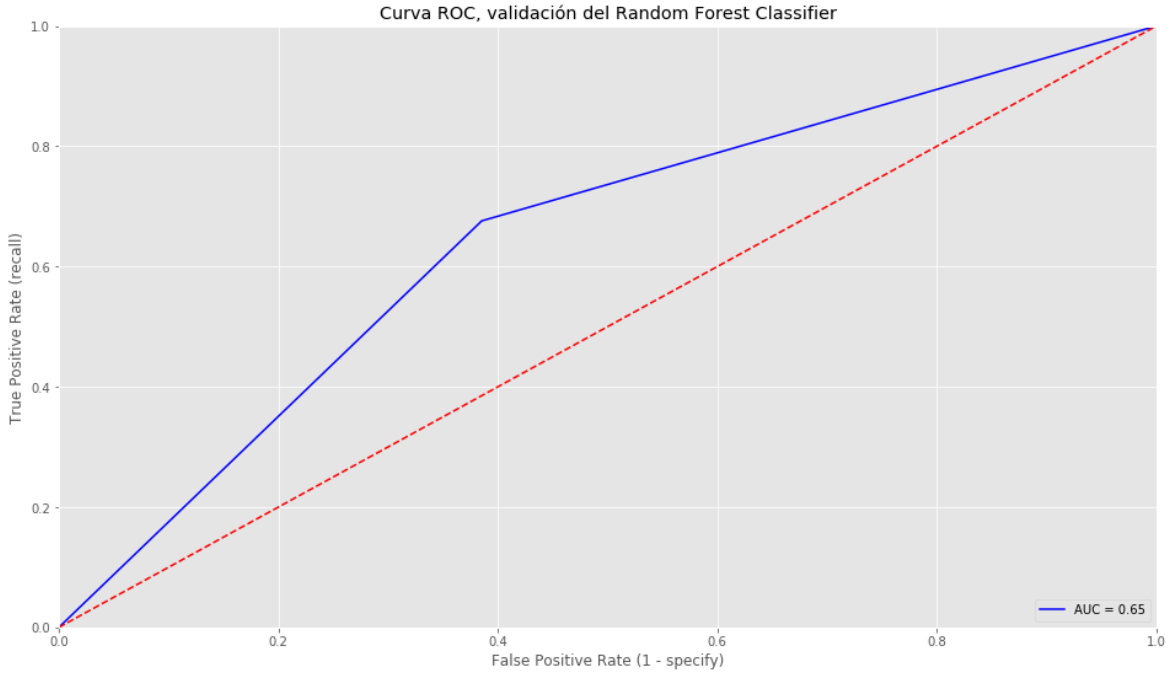

Figura 40 - Curva ROC, validación del Ramdon Forest Classifier

4.5. Evaluación del desarrollo de la metodología

Ahora, que se ha determinado el mejor modelo que presenta mejor calidad desde la perspectiva del análisis de datos, se procede a evaluar y revisar los pasos seguidos en la metodología. El fin de esta fase es decidir la aprobación o no del uso de los resultados del

obtenidos en el análisis, esta evaluación se debe realizar en conjunto con la Dirección del Sistema de Admisión de la UNAH.

Evaluación con resultados del negocio

Como se planteó en la sección 4.1, los objetivos del negocio están muy ligados a determinar los factores que afectan al aspirante al ser o no **admitidos** en la UNAH, así como los aspirantes vulnerables a reprobar la PAA. Bajo esta premisa se ha seleccionado un algoritmo de aprendizaje entre tres (3) posibles, considerando ganador aquel que logre identificar con mejor precisión si un aspirante será **admitido** o no a la UNAH con simplemente inscribirse y sin tener que realizar la PAA.

En la Tabla 25 se observa que la precisión (accuracy) que alcance el modelo final generado con la técnica seleccionada Random Forest Classifier, utilizando un conjunto de datos desconocidos en el entrenamiento es del 66%, mientras que la especificidad para determinar si será **no admitido** es del 61%.

La fiabilidad que expresa el modelo en una comparación del 100% es medianamente alta, con lo que se puede comenzar a generar estrategias sobre este conjunto de aspirantes vulnerables, tal que puedan mejorar su rendimiento en el desarrollo de la Prueba de Aptitud Académica. Además, será necesario evaluar el modelo dentro de su aplicación real y utilizando los datos del sistema de admisión a nivel nacional, ya que solo se han considerado aspirantes que involucrados en la región sur del país.

Revisión del proceso

El proceso de minería de datos se ha realizado de manera minuciosa desde el inicio de la metodología, se ha considerado la exploración, transformación y limpieza de cada campo proporcionado tanto para los datos de la muestra como el de validación, con el objetivo de utilizar en el proceso de selección y validación del modelo la mejor calidad de los datos.

Determinación de próximos pasos

Basado en la evaluación con los resultados del negocio y la revisión del proceso, se considera el modelo final generado lo suficientemente aproximado a las necesidades del negocio planteadas, por lo que se procede a poner el mismo en operación.

4.6. Despliegue

La creación y evaluación del modelo con respecto al negocio, no implica el fin del proyecto. Por lo que es necesario organizar los conocimientos adquiridos para ser presentados de manera utilizable por el negocio. Esto implica poner en operación el modelo generado a disposición de la organización en su toma de decisiones. En otras palabras, que en cada proceso de admisión los encargados de la DSA puedan identificar aspirantes vulnerables a reprobar la PAA y generar estrategias sobre ellos.

Planificación

Para que se obtenga la operabilidad del modelo, es indispensable que se realicen los siguientes pasos de instalación y ejecución:

Instalación de herramientas necesarias

En la Tabla 26 se encuentran las herramientas que se deben instalar para ejecutar sin ningún inconveniente el proyecto completo, en Anaconda se crearon las notebooks utilizadas, las cuales conectan con la base de datos de MySQL con MariaDB que proporciona XAMPP y de la cual el *Dashboard* en PowerBI extrae los datos para la visualización de los datos proporcionados y generados.

Tabla 26 - Herramientas de instalación necesarias

Herramienta	Versión	Enlace
Anaconda	Python 3+	Descargar
XAMPP	7.3.21 / PHP 7.3.21	Descargar
Power BI	Enero 2019	Descargar

Una vez instaladas las herramientas, se deben incorporar librerías de Python necesarias para que las notebooks diseñadas con Jupyter funcionen correctamente, a continuación, se enlistan dichas librerías utilizando el comando pip:

- **pip install geocoder (1.38.1)**: biblioteca de codificación geográfica.
- **pip install plotly (4.6.0)**: librería de visualización de datos interactiva de código abierto.
- **pip install PyMySQL (0.9.3)**: este paquete contiene una biblioteca del cliente MySQL diseñado para ser utilizado con Python
- **pip install xgboost (0.90)**: paquete del algoritmo XGBoost

De no ser posible el instalar las versiones exactas planteadas para cada librería, se debe asegurar estas sean superior, las demás librerías utilizadas vienen integradas en el ambiente de anaconda utilizando Python con versión 3 o superior.

A continuación, se describen los pasos a seguir para la correcta ejecución del proyecto, los mismos se muestran de forma de manera gráfica en el Anexo F – Ejecución del proyecto completo:

1. Iniciar Jupyter Notebook ya sea bajo comando utilizando el Promt de anaconda, o mediante el uso de la interfaz de Anaconda Navigator.
2. Iniciar Apache y MySQL mediante el panel de XAMPP.
3. Ejecución del proyecto en el notebook.
4. Inicio del cuadro de mando diseñado en Power BI.

Monitoreo

La monitorización es una etapa importante dentro de la puesta en producción del modelo generado, por ello es indispensable realizar revisiones de la operabilidad del modelo en cada proceso de admisión con el objetivo de evitar resultados no deseados. Además, realizar ajustes o reentrenar el modelo con los nuevos datos recolectados.

Capítulo 5

5. Diseño del cuadro de mando (*dashboard*)

Para el diseño de un cuadro de mando orientado al nivel operativo, como se planteó en la sección 2.3.3, es necesario diseñar un mapa estratégico con el fin de establecer el conjunto de objetivos a alcanzar. Esta etapa queda resuelta en la fase de la Comprensión del negocio de la sección 4.1, en el desarrollo de la metodología, donde se plantearon los objetivos del negocio y se evaluó el contexto del negocio, determinando recursos y personas involucrados para el desarrollo del proyecto. Basados en los requerimientos en forma de preguntas, se planteó el diseño de un cuadro de mando bajo perspectiva del cliente y procesos internos.

El objetivo general de la investigación está estrechamente relacionado con la detección temprana de los aspirantes vulnerables a reprobar la PAA y dar un seguimiento a los procesos de admisión. Conociendo esta premisa, en la Comprensión de los datos resuelto en la sección 4.2, se determina una agrupación de los datos enfocados en la admisión de los aspirantes a la UNAH. A continuación, se plantean dichos grupos y los *dashboards* elaborados para cada uno de ellos. Cabe mencionar que el año de aplicación de la PAA, centro de aplicación, proceso del año realizado y departamento de nacimiento se consideran como filtros para cada posible gráfica, tal que permita realizar una inspección específica sobre la información a determinar:

Resultados del proceso de la Prueba de Aptitud Académica (PAA)

En la Figura 41 se muestra un *dashboard* con diferentes gráficas que expresan los resultados de la PAA en cada proceso realizado a través de los años. Además, incluye una comparativa sobre la admisión de los aspirantes en su primera, segunda y tercera opción de estudios universitarios.

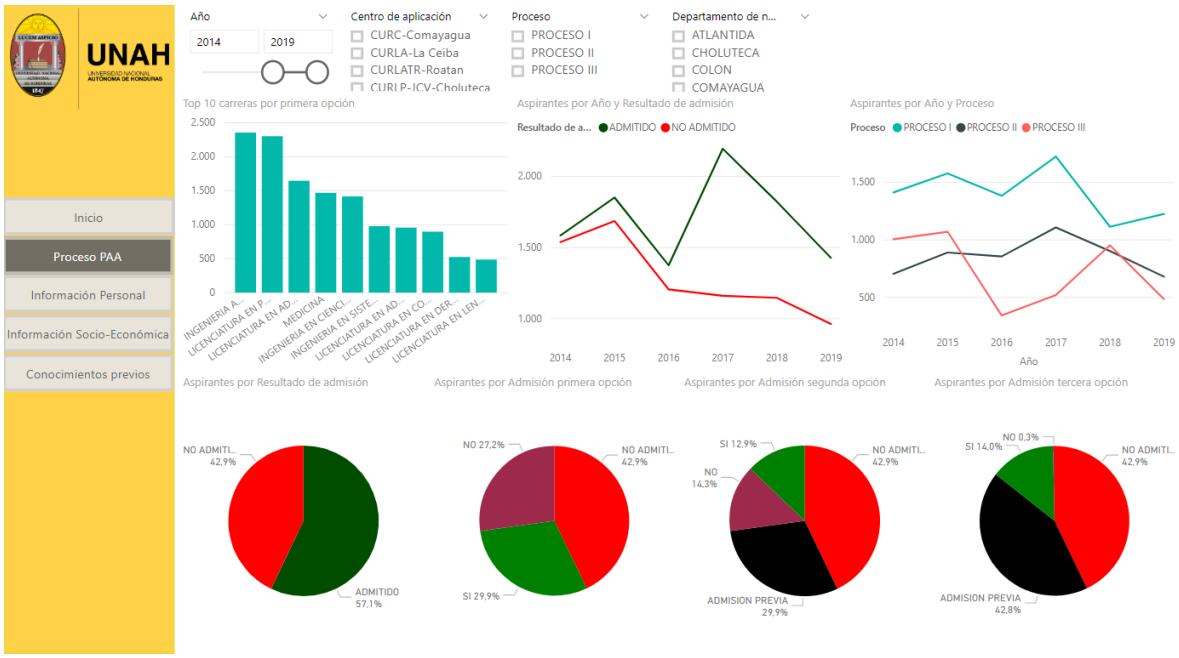

Figura 41 - Cuadro de mando, resultados del proceso PAA

Información personal del aspirante

Se estudian los campos relacionados con la información personal del aspirante como ser el sexo, edad, estado civil, así como la dirección de residencia y nacimiento la determinar una relación de estos campos con el aprobar o reprobar de la PAA por parte de los aspirantes a cursar estudios universitarios en la UNAH, tal y como se muestra en la Figura 42.

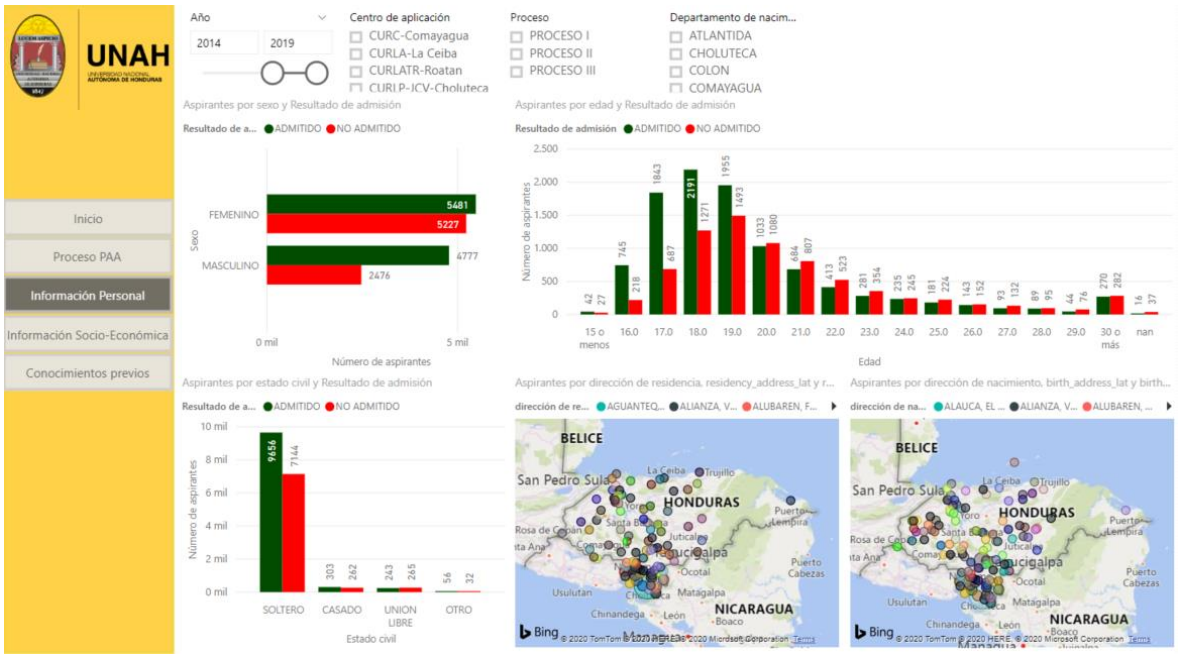

Figura 42 - Cuadro de mando, resultados por información personal del aspirante

Información socioeconómica

En la Figura 43 se muestran los resultados de la PAA relacionados con los campos socioeconómicos que contiene el conjunto de datos proporcionados por la DSA, con ello se diseñaron gráficas que comparan la admisión con respecto a los ingresos mensuales familiares de los aspirantes.

También se adicionan gráficas que describen las jornadas de estudios y de trabajo de los aspirantes, donde rápidamente en forma de ejemplo se observa que en los años del 2014 al 2019, tienden a reprobar aquellos aspirantes con ingresos bajos que estudian por la tarde y trabajan por la noche.

Figura 43 - Cuadro de mando, resultados por información socioeconómica

Conocimientos académicos previos

En cuanto a los conocimientos académicos previos, se plantean gráficas que relacionan la admisión del aspirante según el instituto de procedencia, el nivel de educación alcanzado hasta ese momento y la facultad de la carrera de estudios universitarios de primera opción, así como el intervalo de tiempo (lapso) que toman los aspirantes una vez graduados de educación media para iniciar sus estudios universitarios, tal y como se muestra en la Figura 44.

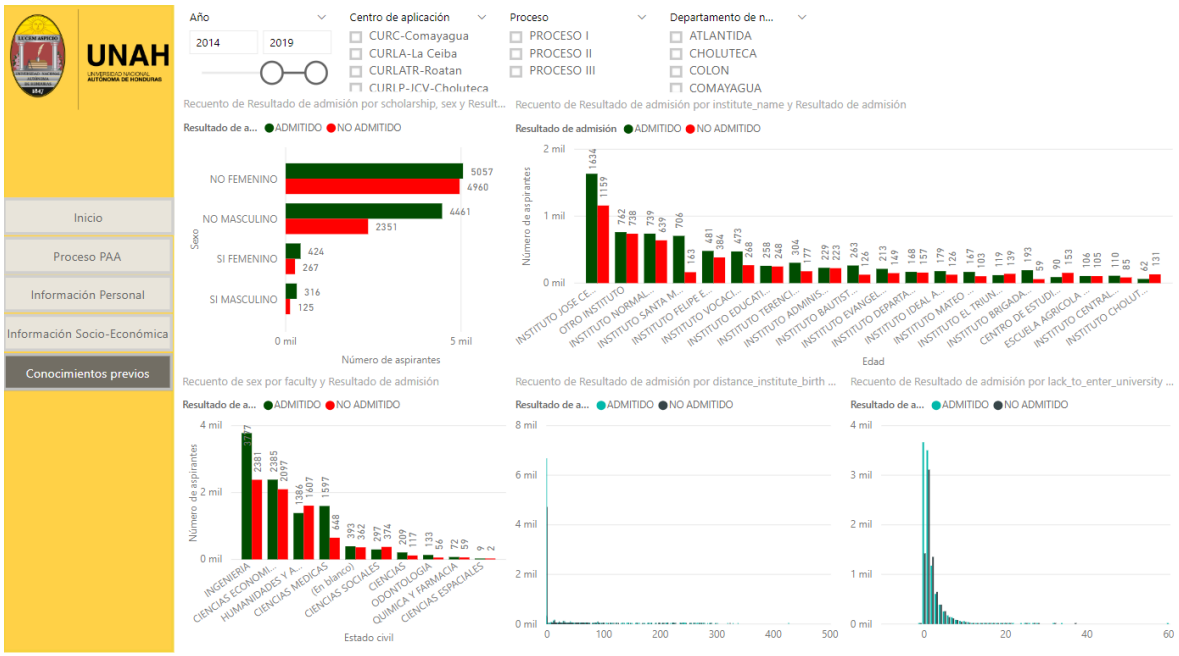

Figura 44 - Cuadro de mando, resultados por conocimientos de estudios previos

Capítulo 6

6. Conclusiones y trabajo futuro

6.1. Conclusiones

Se ha implementado exitosamente la combinación entre las metodologías CRISP-DM y SEMMA, la cual ha permitido comprender las necesidades tanto de la Dirección del Sistema de Admisión (DSA) como de la Universidad Nacional Autónoma de Honduras (UNAH), en cuanto al mejoramiento del rendimiento académico de los aspirantes que se someten a la Prueba de Aptitud Académica (PAA), como de los estudiantes durante sus estudios universitarios. Estas metodologías han enfocado su estudio en realizar una detección temprana de aspirantes vulnerables a reprobar la PAA, en primera instancia utilizando los datos recopilados en la región sur del país.

La comprensión y preparación de los datos ha permitido obtener un conjunto de datos representativo y de calidad para generar un modelo de aprendizaje automático supervisado de clasificación basado en el algoritmo Random Forest Classifier que proporciona la librería *sklearn*, éste es capaz de determinar en un 66% de precisión si un aspirante aprobará o no la Prueba de Aptitud Académica. Además, este modelo es sensible en un 61% a determinar los aspirantes que obtendrán un **no admitido** una vez publicados los resultados de la PAA. El porcentaje de especificidad puede considerarse relativamente bajo. Sin embargo, es un punto de partida para generar nuevas estrategias donde se aglomeren los aspirantes vulnerables, esto a pesar de contar con ciertos falsos negativos.

La comprensión del negocio ha permitido determinar los recursos y personas involucradas en el desarrollo del proyecto, así como los requerimientos en forma de preguntas, con los cuales se ha diseñado un cuadro de mando dinámico bajo la perspectiva

del cliente y procesos internos enfocado en la admisión de los aspirantes a la UNAH. Éste separa los *dashboards* en grupos según los datos proporcionados, como ser: resultados del proceso de la Prueba de Aptitud Académica (PAA), resultados por información personal del aspirante, resultados según situación socioeconómica y resultados basados en los conocimientos de estudios previos.

6.2. Trabajo futuro

Hasta el momento, se ha enfocado el análisis de los datos para un caso puntual "detección temprana de aspirantes a reprobar la PAA". Sin embargo, durante la etapa de comprensión de los datos, se ha visto que la información recopilada mediante la aplicación de la PAA tiene mucho más potencial. A continuación, se listan algunas de las posibilidades a realizar con los datos:

- Análisis del impacto que ocasionan factores externos (p. ej. protestas o COVID19) en el número de aspirantes que se someten al proceso de admisión a la UNAH.
- Influencia de los estudios de educación media en la selección y admisión de la primera opción de carrera universitaria, recordando que ser **admitido** en la UNAH no implica realizar estudios en una carrera universitaria de específica, dado que estas tienen sus propios criterios de aprobación.
- Adición de reglas de asociación para el descubrimiento de nuevas relaciones en los datos del sistema de admisión.

En cuanto al modelado de los datos, buscar alternativas que permitan utilizar campos categóricos que fueron proporcionados, creados o investigados y han sido descartados por no permitir generalizar el modelo y agregar complejidad al procesamiento (p. ej. el nombre del instituto, carrera de estudios, facultad y modalidad) e intentar aportar valor al modelo final creado, para obtener una mejor precisión en sus resultados.

Además, se debe evaluar el modelo generado con los datos del proceso de admisión a nivel nacional para describir su potencial predictivo y considerar mejoras que permitan su utilización en toda la universidad independientemente de la región del país.

Bibliografía

[1] S. M. Van Lamoen, “La Responsabilidad Social Universitaria,” *Synerg. Chili*, vol. 9, pp. 35–49, 2013, doi: 10.18041/2382-3240/saber.2010v5n1.2543.

[2] “Misión y visión de la UNAH.” https://www.unah.edu.hn/sobre-la-unah/mision-y-vision (accessed Sep. 05, 2020).

[3] Dirección del Sistema de Admisión, “Sistematización del Sistema de Admisión: Buenas Prácticas para una Educación Sin Fronteras,” 2015.

[4] Javier Solorzano, “Por qué sí debe haber examen de admisión a las universidades públicas - YouTube,” Oct. 01, 2019. https://www.youtube.com/watch?v=o_2nPrDmYhE&ab_channel=UnoTV (accessed Sep. 05, 2020).

[5] “myKlovr: Virtual College Counselor | AI Based Platform.” https://myklovr.com/ (accessed Sep. 05, 2020).

[6] “Machine learning para acceder a la Universidad | Avances AI.” https://www.avances.ai/machine-learning-admision-universidad/ (accessed Sep. 05, 2020).

[7] “Historia de la UNAH.” https://www.unah.edu.hn/sobre-la-unah/historia (accessed Jun. 18, 2020).

[8] “Centros regionales.” https://www.unah.edu.hn/centros-regionales (accessed Jun. 18, 2020).

[9] “CRAED.” https://sed.unah.edu.hn/craed (accessed Jun. 18, 2020).

[10] “Telecentros UNAH.” https://die.unah.edu.hn/servicios-educativos/telecentros-unah (accessed Jun. 18, 2020).

[11] “Oferta académica.” https://www.unah.edu.hn/oferta-academica (accessed Jun. 18, 2020).

[12] “Graduados.” https://estadistica.unah.edu.hn/sistema-estadistico/graduados (accessed Jun. 18, 2020).

[13] “Organigrama general de la UNAH.” https://www.unah.edu.hn/sobre-la-unah/organigrama (accessed Jun. 18, 2020).

[14] “Atribuciones VRA.” https://vra.unah.edu.hn/sobre-nosotros/atribuciones-vra (accessed Jun. 18, 2020).

[15] “Direcciones Académicas.” https://vra.unah.edu.hn/sobre-nosotros/direcciones-academicas (accessed Jun. 18, 2020).

[16] “Sobre Nosotros, Dirección del Sistema de Admisión.” https://admisiones.unah.edu.hn/sistema-de-admision/sobre-nosotros (accessed Jun. 18, 2020).

[17] Universidad Nacional Autónoma de Honduras, “Normas Académicas de la Universidad Nacional Autónoma de Honduras,” no. 6, p. 105, 2009, [Online]. Available: https://www.unah.edu.hn/sobre-la-unah/normas-academicas.

[18] “Reglamento de Admisión de los estudiantes a la UNAH.” Tegucigalpa, 2008, [Online]. Available: https://www.unah.edu.hn/estudiantes/admisiones.

[19] “Inicio - College Board.” https://latam.collegeboard.org/ (accessed Jul. 12, 2020).

[20] “En Honduras seis de cada cien personas hablan el idioma inglés - Diario El Heraldo.” https://www.elheraldo.hn/pais/1096086-466/en-honduras-seis-de-cada-cien-personas-hablan-el-idioma-inglés (accessed Jul. 12, 2020).

[21] “Resultados PAA - College Board.” https://latam.collegeboard.org/resultados/resultados-paa/ (accessed Jul. 12, 2020).

[22] “El proceso de análisis de datos - Marketing Analítico.” https://www.marketing-analitico.com/analitica-web/proceso-analisis-datos/ (accessed Jul. 19, 2020).

[23] “Análisis descriptivo de datos en educación - Fco. Javier Tejedor Tejedor - Google

Libros.” https://books.google.es/books?hl=es&lr=&id=trlCB7wtTcMC&oi=fnd&pg=PP2&dq=analisis+descriptivo+de+datos&ots=BwW3nKNkyU&sig=0rq6O3dYNPMg7wvLCtlMXiCFmNI#v=onepage&q=analisis descriptivo de datos&f=false (accessed Jul. 19, 2020).

[24] I. N. de Estadística, “Tipos de gráficas.”

[25] “Diagrama de caja - Wikipedia, la enciclopedia libre.” https://es.wikipedia.org/wiki/Diagrama_de_caja (accessed Sep. 03, 2020).

[26] “BBC Bitesize - GCSE Maths - Representing data - Edexcel - Revision 7,” *BBC Bitesize*, Accessed: Sep. 03, 2020. [Online]. Available: https://www.bbc.com/bitesize/guides/zc7sb82/revision/7.

[27] “3 tipos de análisis de datos para mejorar la toma de decisiones.” https://www.pragma.com.co/blog/3-tipos-de-analisis-de-datos-para-mejorar-la-toma-de-decisiones (accessed Jul. 19, 2020).

[28] “¿Qué es el Análisis Predictivo? - Big Data Social.” http://www.bigdata-social.com/que-es-el-analisis-predictivo/ (accessed Sep. 03, 2020).

[29] “Análisis predictivo: Tres cosas que es necesario saber - MATLAB & Simulink.” https://es.mathworks.com/discovery/predictive-analytics.html (accessed Sep. 03, 2020).

[30] “¿Qué es Machine Learning? – Cleverdata.” https://cleverdata.io/que-es-machine-learning-big-data/ (accessed Jul. 19, 2020).

[31] M. H. Tom M. Mitchell, “CHAPTER 14 Key Ideas in Machine Learning,” 2017. [Online]. Available: www.cs.cmu.edu/~tom/mlbook.html.

[32] “Algunos conceptos básicos detrás de el análisis Machine Learning GeoSapience.” https://www.geo-sapience.com/algunos-conceptos-machine-learning/ (accessed Sep. 02, 2020).

[33] “Técnicas y Algoritmos de Machine Learning: qué tipo de problemas se pueden resolver usando Machine Learning. – Modeling reality, generating software.” https://genexus.blog/es_ES/artificial-intelligence/tecnicas-y-algoritmos-de-machine-learning/ (accessed Jul. 19, 2020).

[34] L. Bennett, “Machine Learning in ArcGIS,” Nov. 27, 2017. https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/machine-learning-in-arcgis/ (accessed Sep. 02, 2020).

[35] “Diferencia entre algoritmos de clasificación y regresión - Ligdi González.” https://ligdigonzalez.com/diferencia-entre-algoritmos-de-clasificacion-y-regresion/ (accessed Sep. 03, 2020).

[36] “Evaluación de modelos de ML - Amazon Machine Learning.” https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/evaluating_models.html (accessed Sep. 03, 2020).

[37] AIREL PEREZ SUÁREZ, “Algoritmos de Agrupamiento para la Generación de Cubrimientos en Colecciones de Documentos,” 2008.

[38] “Machine Learning: Seleccion Metricas de clasificacion - sitiobigdata.com.” https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/# (accessed Sep. 03, 2020).

[39] “Selección de métricas para los modelos de aprendizaje automático | Fayrix.” https://fayrix.com/machine-learning-metrics_es (accessed Sep. 03, 2020).

[40] “Evaluation Metrics Machine Learning.” https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ (accessed Sep. 03, 2020).

[41] “Metrics To Evaluate Machine Learning Algorithms in Python.” https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/ (accessed Sep. 03, 2020).

[42] “Aprendizaje automático: como seleccionar métricas de regresión - sitiobigdata.com.” https://sitiobigdata.com/2019/05/27/aprendizaje-automatico-seleccionando-metricas-regresion/ (accessed Sep. 03, 2020).

[43] “Cómo implementar correctamente un Cuadro de Mando Integral - PrevenBlog.” https://prevenblog.com/como-implementar-un-cuadro-de-mando-integral/ (accessed Sep. 04, 2020).

[44] J. L. García Suárez, “Metodología de diseño de un cuadro de mando.,” p. 43, 2010, [Online]. Available:

https://www.unioviedo.es/cecodet/MDL08/docum/met_diseno_cuadro_mando.pdf.

[45] “En los últimos 10 años cobertura universitaria aumentó 1.7% - Diario La Prensa.” https://www.laprensa.hn/honduras/1110688-410/honduras-cobertura-universitaria-educacion- (accessed Sep. 06, 2020).

[46] C. L. Hernández G and M. X. Dueñas R, “Hacia una metodología de gestión del conocimiento basada en minería de datos,” p. 17, 2009, [Online]. Available: https://goo.gl/1eXxSn.

[47] “Metodologías de Data Mining - Auribox Training.” https://blog.auriboxtraining.com/business-intelligence/metodologias-de-data-mining/ (accessed Sep. 04, 2020).

[48] “EL BUHO ANALÍTICO: METODOLOGÍA SEMMA.” http://elbuhoanaltico.blogspot.com/2012/02/metodologia-semma.html (accessed Jul. 09, 2020).

[49] “Anaconda | Individual Edition.” https://www.anaconda.com/products/individual (accessed Aug. 26, 2020).

[50] “Infografía del estado del ecosistema del desarrollador en 2020 | JetBrains: Developer Tools for Professionals and Teams.” https://www.jetbrains.com/es-es/lp/devecosystem-2020/ (accessed Aug. 26, 2020).

[51] “ Python vs Scala | Know The Top 9 Significance Differences.” https://www.educba.com/python-vs-scala/ (accessed Aug. 26, 2020).

[52] “Project Jupyter | Home.” https://jupyter.org/ (accessed Aug. 26, 2020).

[53] “XAMPP Installers and Downloads for Apache Friends.” https://www.apachefriends.org/es/index.html (accessed Aug. 26, 2020).

[54] “MySQL Database Service | Oracle.” https://www.oracle.com/mysql/ (accessed Aug. 26, 2020).

[55] “Power BI: la herramienta de Business Intelligence de Office 365.” https://www.nunsys.com/power-bi/ (accessed Aug. 26, 2020).

[56] “COMUNICADO 20 DE SEPTIEMBRE DE 2016 - Blogs UNAH.” https://blogs.unah.edu.hn/dircom/comunicado-20-de-septiembre-de-2016 (accessed

Aug. 11, 2020).

Anexos

Anexo A – Ingeniería de características sobre el conjunto de datos

Se describen en la Tabla A 1 las acciones realizadas sobre los campos proporcionados como parte de la preparación de los datos en el desarrollo de la metodología.

Tabla A 1 - Ingeniería de características sobre el conjunto de datos

Campo	Acción	Descripción
test_application_date	Reemplazar por test_application_process	Identificar en que proceso está realizando el aspirante la PAA.
birth_department, birth_municipality, birth_country	Reemplazar por birth_address_lat, birth_address_lng	Obtener la longitud y latitud para la dirección de nacimiento.
residency_department, residency_municipality, residency_country	Reemplazar por residency_address_lat, residency_address_lng	Obtener la longitud y latitud para la dirección de residencia.
institute_country, institute_department, institute_municipality	Reemplazar por institute_address_lat, institute_address_lng	Obtener la longitud y latitud para la dirección del instituto.
birth_year	Agregar aproximated_age	Identificar la edad aproximada del aspirante al momento de someterse al proceso de admisión.
graduation_year	Agregar lack_to_enter_university	Conocer el intervalo de tiempo entre graduación y año de realización de la PAA.
study_center, application_center	Reemplazar por Study_center_application	Determinar si el centro regional universitario es el

Campo	Acción	Descripción
		mismo donde se inscribió para realizar la PAA.
option_one, option_two, option_three	Reemplazar paa_score_option_x	Este campo se reemplaza por su valor requerido para ser admitido en la carrera.
pccns_rate, pam_rate	Agregar pam_option_x pccns_option_x compete_option_x	Identificar si el aspirante debe realizar o no una prueba extra o competir por una plaza.
admission_option	Reemplazar por admitted_option_one, admitted_option_two, admitted_option_three	Determinar si fue admitido a la primera, segunda o tercera opción de matrícula
languages	Transformar	Identificar si el aspirante habla uno o más idiomas y luego indexar
sex, bilingual, employee, scholarship, live_with_parents, separated_parents, languages, test_application_process, pam_option_x, pccns_option_x, compete_option_x, monthly_income, computing_level, education_level, study_day, father_education_level, mother_education_level	Transformar	Indexar los campos

Anexo B - Característica de los datos proporcionados

En la Tabla A 2 se describe el tipo de los datos proporcionados por la Dirección del Sistema de Admisión, así como el nuevo nombre asignado a cada campo para identificarlo más fácilmente al momento de tratarlos.

Tabla A 2 - Característica de los datos proporcionados

N°	Variable	Nuevo Nombre	Tipo de dato
1	Fecha de aplicación Pruebas	test_application_date	fecha
2	Codigo Numero Solicitud	request_number_code	numérico
3	Codigo Identidad	identity_code	numérico
4	Sexo	sex	texto
5	Estado civil	marital_status	texto
6	Departamento Nacimiento	birth_department	texto
7	Municipio nacimiento	birth_municipality	texto
8	Pais nacimiento	birth_country	texto
9	Departamento Residencia	residency_department	texto
10	Municipio Residencia	residency_municipality	texto
11	Pais Residencia	residency_country	texto
12	Nacionalidad	nationality	texto
13	Año Nacimiento	birth_year	numérico
14	Pais Instituto	institute_country	texto
15	Departamento del instituto	institute_department	texto
16	Municipio instituto	institute_municipality	texto
17	Nombre del Instituto	institute_name	texto
18	Sector del Instituto	institute_sector	texto
19	bilingüe	bilingual	texto
20	Año Graduacion	graduation_year	numérico
21	Carrera de educacion Media	meddle_school_career	texto
22	Centro estudios	study_center	texto
23	Primera carrera	option_one	texto
24	Segunda carrera	option_two	texto
25	Tercera Carrera	option_three	texto

N°	Variable	Nuevo Nombre	Tipo de dato
26	Trabaja	employee	texto
27	Jornada Trabajo	working_day	texto
28	Area Trabajo	work_area	texto
29	Becado	scholarship	texto
30	Centro PAA	application_center	texto
31	INDICE ADMISION	admision_rate	numérico
32	PAA verbal	verbal_paa_score	numérico
33	PAA Matematica	math_paa_score	numérico
34	PCCNS indice	pccns_rate	numérico
35	PAM indice	pam_rate	numérico
36	Admision	admitted	texto
37	CARRERA ADMISION	admission_option	texto
38	Numero hermanos	number_of_brothers	numérico
39	Numero Hijos	number_of_children	numérico
40	Vive Padres	live_with_parents	texto
41	Padres Separados	separated_parents	texto
42	Ingreso mendual	monthly_income	texto
43	Nivel computacional	computing_level	texto
44	Nivel Educativo	education_level	texto
45	Grupo etnico	ethnic_group	texto
46	Jornada estudio	study_day	texto
47	Idiomas	languages	texto
48	Nivel educativo padre	father_education_level	texto
49	Nivel Educativo Madre	mother_education_level	texto
50	Sector Trabajo Madre	mother_job_sector	texto
51	Sector Trabajo Padre	father_job_sector	texto
52	Razon No trabaja madre	why_mother_doesnt_work	texto
53	razon no trabaja padre	why_father_doesnt_work	texto
54	periodo matricula	enrollment_period	texto

Anexo C – Oferta académica

Se observa en la Tabla A 3 información respecto al puntaje mínimo requerido para ser admitido en las diferentes carreras universitarias ofertadas por la Universidad Nacional Autónoma de Honduras. Además, si los aspirantes deben realizar una prueba extra o si deben competir por un plaza.

Tabla A 3 - Oferta académica

Carreras	Puntaje	PAM	PCCNS	Compite
Medicina	1100	NO	SI	SI
Ingeniería Civil	1000	SI	NO	SI
Ingeniería Eléctrica Industrial	1000	SI	NO	SI
Ingeniería en Sistemas	1000	SI	NO	SI
Ingeniería Industrial	1000	SI	NO	SI
Ingeniería Mecánica Industrial	1000	SI	NO	SI
Ingeniería Química	1000	SI	NO	SI
Ingeniería Química Industrial	1000	SI	NO	SI
Odontología	1000	NO	SI	SI
Microbiología	950	NO	SI	SI
Licenciatura en Administración de Empresas	900	NO	NO	NO
Licenciatura en Derecho	900	NO	NO	NO
Enfermería	900	NO	SI	SI
Licenciatura en Fonoaudiología	900	NO	NO	NO
Licenciatura en Antropología	900	NO	NO	NO
Licenciatura en Psicología	900	NO	NO	NO
Arquitectura	900	NO	NO	SI
Química y Farmacia	850	NO	SI	SI
Ingeniería Agroindustrial	827	NO	NO	NO
Ingeniería Agronómica	827	NO	NO	NO
Ingeniería en Ciencias Acuícolas Y Recursos Marinos Costeros	827	NO	NO	NO
Ingeniería Forestal	827	NO	NO	NO

Carreras	Puntaje	PAM	PCCNS	Compite
Licenciatura En Len. Ext. Con Orie. Ingl. Y Frances	800	NO	NO	SI
Licenciatura en Biología	700	NO	NO	NO
Licenciatura en Física	700	NO	NO	NO
Licenciatura en Geología	700	NO	NO	NO
Licenciatura en Matemáticas	700	NO	NO	NO
Técnico Universitario En Metalurgia	700	NO	NO	NO
Licenciatura en Administración Aduanera	700	NO	NO	NO
Licenciatura en Administración de Banca y Finanzas	700	NO	NO	NO
Licenciatura en Administración Pública	700	NO	NO	NO
Licenciatura En Admon. de Empresas Agropecuarias	700	NO	NO	NO
Licenciatura En Com. Internacional Con Orient. En Agroind.	700	NO	NO	NO
Licenciatura en Comercio Internacional	700	NO	NO	NO
Licenciatura en Contaduría Pública y Finanzas	700	NO	NO	NO
Licenciatura en Economía	700	NO	NO	NO
Licenciatura en Economía Agrícola	700	NO	NO	NO
Licenciatura en Ecoturismo	700	NO	NO	NO
Licenciatura en Informática Administrativa	700	NO	NO	NO
Licenciatura en Mercadotecnia	700	NO	NO	NO
Técnico Universitario en Administración de Empresas Cafetaleras	700	NO	NO	NO
Técnico Universitario en Alimentos y Bebidas	700	NO	NO	NO
Técnico Universitario en Microfinanzas	700	NO	NO	NO
Técnico Universitario en Tecnología de Alimentos	700	NO	NO	NO
Licenciatura en Astronomía y Astrofísica	700	NO	NO	NO
Licenciatura en Ciencia y Tecnologías de la Información Geográfica	700	NO	NO	NO

Carreras	Puntaje	PAM	PCCNS	Compite
Técnico Universitario en Sistemas de Información Geografía con Énfasis en Catastro	700	NO	NO	NO
Nutrición	700	NO	SI	SI
Técnico Universitario En Radio tecnologías (Radiología e Imágenes)	700	NO	NO	NO
Técnico Universitario en Terapia Funcional	700	NO	NO	NO
Licenciatura en Desarrollo Local	700	NO	NO	NO
Licenciatura en Historia	700	NO	NO	NO
Licenciatura en Periodismo	700	NO	NO	NO
Licenciatura en Sociología	700	NO	NO	NO
Licenciatura en Trabajo Social	700	NO	NO	NO
Técnico Universitario en Desarrollo Municipal	700	NO	NO	NO
Técnico Universitario en Lengua de Señas	700	NO	NO	NO
Licenciatura en Educación Física	700	NO	NO	NO
Licenciatura en Filosofía	700	NO	NO	NO
Licenciatura en Letras	700	NO	NO	NO
Licenciatura en Música	700	NO	NO	NO
Licenciatura en Pedagogía	700	NO	NO	NO
Técnico Universitario en Educación Básica para la Enseñanza del Español	700	NO	NO	NO
Licenciatura en Química Industrial	700	NO	NO	NO
Técnico Universitario en Calidad del Café	700	NO	NO	NO
Técnico Universitario en Producción Agrícola	700	NO	NO	NO

Anexo D – Validación cruzada Random Forest Classifier

En la Tabla A 4 se observan las diferentes iteraciones generadas por la validación cruzada utilizando el algoritmo Random Forest Classifier, en esta se varía el criterio de partición, el número de estimadores y la profundidad de los árboles.

Tabla A 4 - Validación cruzada Random Forest Classifier

	criterion	max_depth	n_estimators	mean_test_score	rank_test_score
1	gini	5	70	64.19	45
2	gini	5	75	63.96	49
3	gini	5	80	64.30	42
4	gini	5	85	64.23	43
5	gini	5	90	64.30	41
6	gini	10	70	65.63	12
7	gini	10	75	65.65	11
8	gini	10	80	65.53	21
9	gini	10	85	65.45	23
10	gini	10	90	65.59	16
11	gini	15	70	65.38	32
12	gini	15	75	65.74	5
13	gini	15	80	65.44	24
14	gini	15	85	65.69	6
15	gini	15	90	65.89	2
16	gini	20	70	65.13	39
17	gini	20	75	65.47	22
18	gini	20	80	65.38	31
19	gini	20	85	65.04	40
20	gini	20	90	65.31	35

	criterion	max_depth	n_estimators	mean_test_score	rank_test_score
21	gini	25	70	65.44	26
22	gini	25	75	65.33	34
23	gini	25	80	65.43	27
24	gini	25	85	65.55	19
25	gini	25	90	65.19	37
26	entropy	5	70	64.20	44
27	entropy	5	75	63.90	50
28	entropy	5	80	64.04	48
29	entropy	5	85	64.18	46
30	entropy	5	90	64.16	47
31	entropy	10	70	65.35	33
32	entropy	10	75	65.57	17
33	entropy	10	80	65.56	18
34	entropy	10	85	65.44	25
35	entropy	10	90	65.69	7
36	entropy	15	70	65.67	9
37	entropy	15	75	65.84	4
38	entropy	15	80	65.91	1
39	entropy	15	85	65.60	14
40	entropy	15	90	65.87	3
41	entropy	20	70	65.63	13
42	entropy	20	75	65.42	28
43	entropy	20	80	65.68	8
44	entropy	20	85	65.60	15
45	entropy	20	90	65.67	10

	criterion	max_depth	n_estimators	mean_test_score	rank_test_score
46	entropy	25	70	65.14	38
47	entropy	25	75	65.40	30
48	entropy	25	80	65.19	36
49	entropy	25	85	65.41	29
50	entropy	25	90	65.53	20

Anexo E – Validación cruzada XGBoost Classifier

En la Tabla A 5 se observan las diferentes iteraciones generadas por la validación cruzada utilizando el algoritmo XGBoost Classifier, en esta se varía el criterio de partición, el número de estimadores y la profundidad de los árboles.

Tabla A 5 - Validación cruzada XGBoost Classifier

	criterion	max_depth	n_estimators	mean_test_score	rank_test_score
1	gini	5	70	66.04	3
2	gini	5	75	66.07	1
3	gini	5	80	66.00	5
4	gini	5	85	65.94	7
5	gini	5	90	65.92	9
6	gini	10	70	65.53	11
7	gini	10	75	65.43	15
8	gini	10	80	65.29	19
9	gini	10	85	65.41	17
10	gini	10	90	65.49	13
11	gini	15	70	64.81	27
12	gini	15	75	64.92	21
13	gini	15	80	64.89	23
14	gini	15	85	64.82	25
15	gini	15	90	64.73	29
16	gini	20	70	64.39	49
17	gini	20	75	64.52	41
18	gini	20	80	64.45	43
19	gini	20	85	64.44	47
20	gini	20	90	64.45	45

	criterion	max_depth	n_estimators	mean_test_score	rank_test_score
21	gini	25	70	64.73	31
22	gini	25	75	64.67	35
23	gini	25	80	64.57	39
24	gini	25	85	64.58	37
25	gini	25	90	64.72	33
26	entropy	5	70	66.04	3
27	entropy	5	75	66.07	1
28	entropy	5	80	66.00	5
29	entropy	5	85	65.94	7
30	entropy	5	90	65.92	9
31	entropy	10	70	65.53	11
32	entropy	10	75	65.43	15
33	entropy	10	80	65.29	19
34	entropy	10	85	65.41	17
35	entropy	10	90	65.49	13
36	entropy	15	70	64.81	27
37	entropy	15	75	64.92	21
38	entropy	15	80	64.89	23
39	entropy	15	85	64.82	25
40	entropy	15	90	64.73	29
41	entropy	20	70	64.39	49
42	entropy	20	75	64.52	41
43	entropy	20	80	64.45	43
44	entropy	20	85	64.44	47
45	entropy	20	90	64.45	45

	criterion	max_depth	n_estimators	mean_test_score	rank_test_score
46	entropy	25	70	64.73	31
47	entropy	25	75	64.67	35
48	entropy	25	80	64.57	39
49	entropy	25	85	64.58	37
50	entropy	25	90	64.72	33

Anexo F – Ejecución del proyecto completo

En la Figura A 1 se muestra el comando utilizando en el promt de anaconada para iniciar jupyter notebook, herramienta necesaria para la codificación del proyecto.

Iniciar Jupyter Notebook

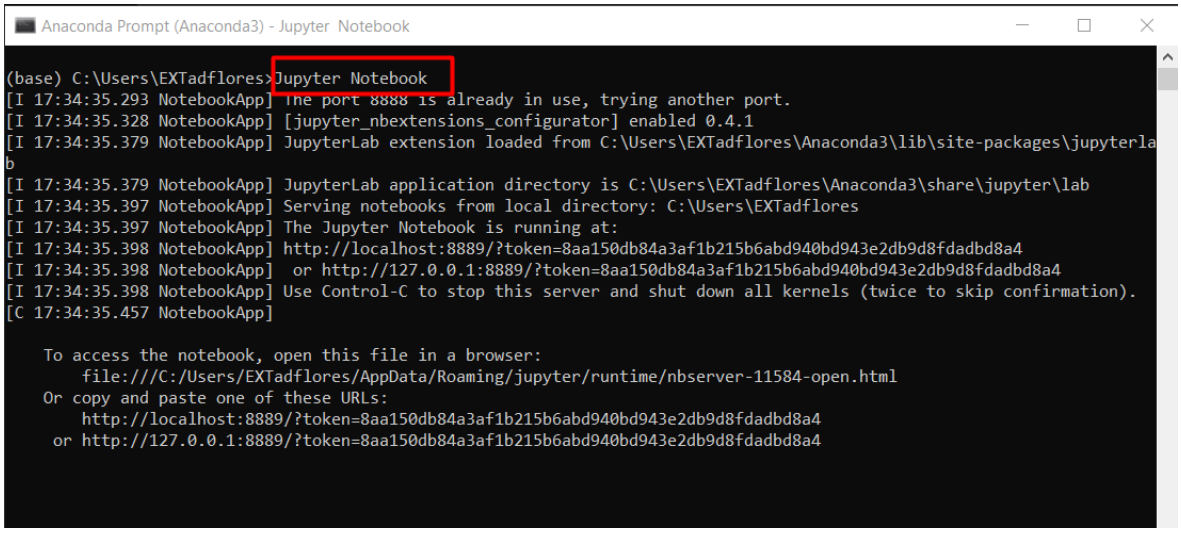

Figura A 1 - Iniciar Jupyter Notebook bajo comando en el Prompt

Además del promt de anaconda, se puede utilizar el Anaconda Navigator como se muestra en la Figura A 2.

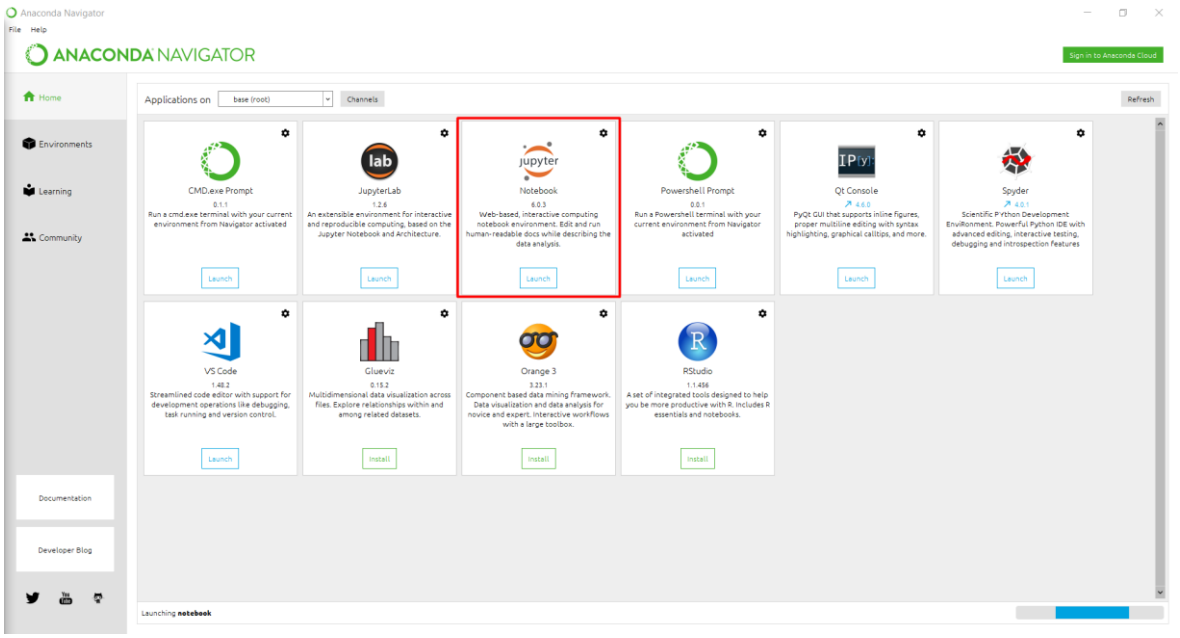

Figura A 2 - Ejecución de Jupyter Notebook mediante interfaz Anaconda Navigator

Iniciar Apache y MySQL

Es necesario tener activado el servicio de apache para establecer la conexión entre el servidor de base de datos MySQL y el jupyter, tal y como se muestra en la Figura A 3.

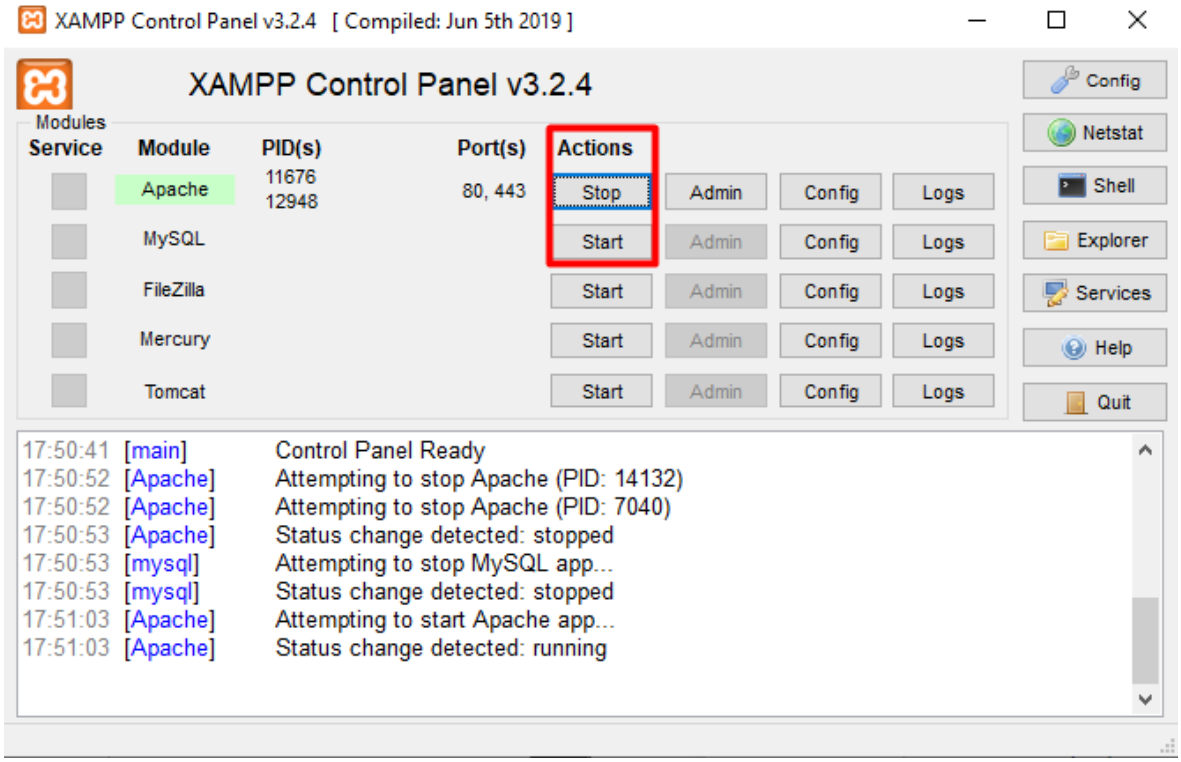

Figura A 3 - Ejecución de Apache y MySQL

Puesta en marcha del proyecto

Una vez se tenga iniciado el servicio de base de datos y jupyter notebook, se reinicia y ejecuta todo el proyecto del notebook tal y como se observa en la Figura A 4.

Cuadro de mando (dashboard)

En la Figura A 5 se observa la pantalla principal del cuadro mando generado como herramienta de apoyo a las estrategias que genera la Dirección del Sistema de Admisión de la Universidad Nacional Autónoma de Honduras.

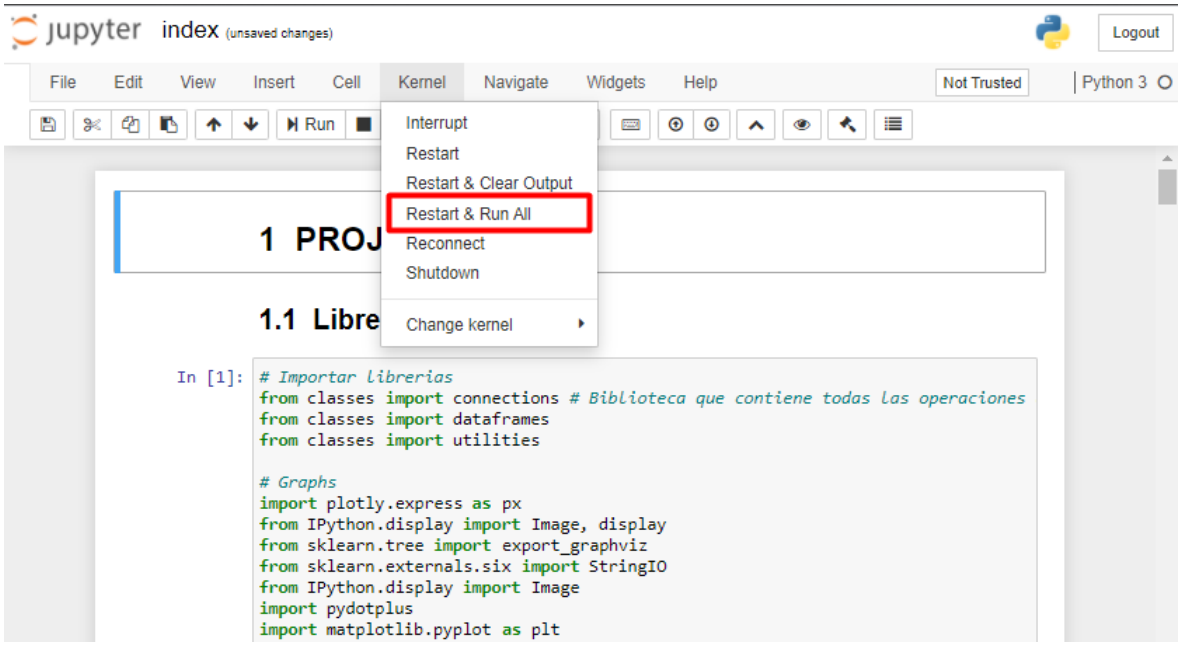

Figura A 4 - Ejecución del proyecto completo en el notebook

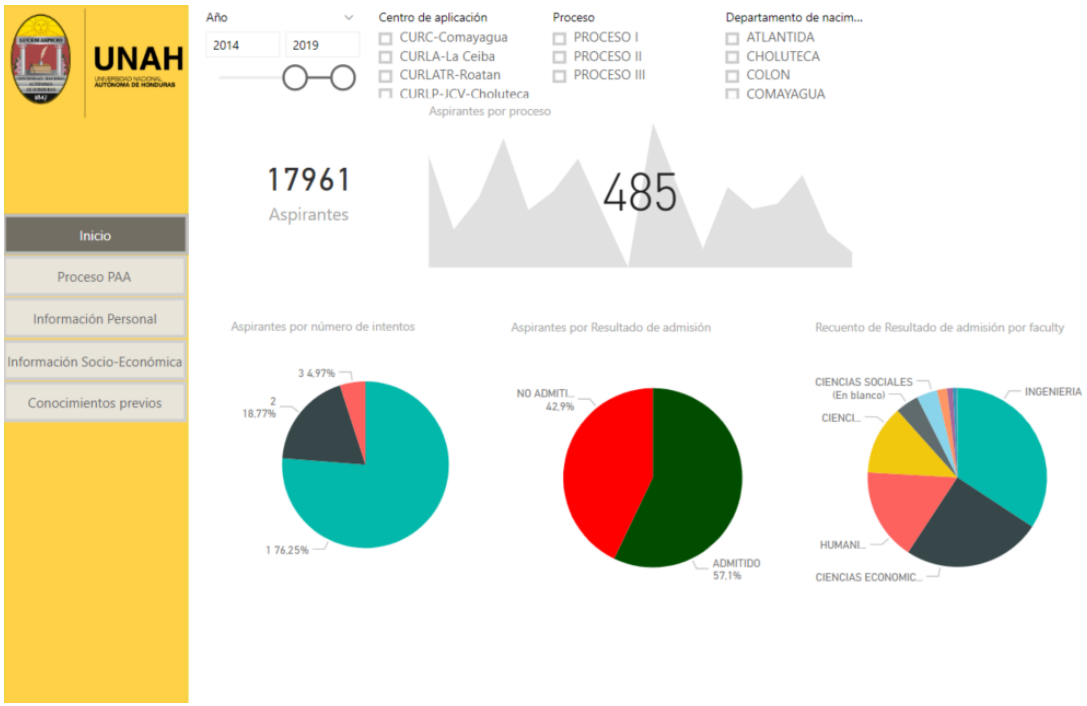

Figura A 5 - Cuadro de mando diseñado en Power BI