PONTIFICIA

UNIVERSIDAD CATOLICA DEL ECUADOR

MAESTRIA EN SISTEMAS DE INFORMACIÓN MENCIÓN CIENCIA DE DATOS

TEMA:

DETERMINAR EL MEJOR ALGORITMO DE CIENCIA DE DATOS PARA UNA PLANIFICACION ACADEMICA AUTOMATIZADA PARA LAS IES DEL ECUADOR

AUTOR: HECTOR HERNANDEZ

TUTOR: Ph.D. JHONNY PINCAY NIEVES

QUITO – ECUADOR

2023

ÍNDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

Después de la pandemia las instituciones de educación superior apoyados por los órganos de control del Ecuador, han adoptado diferentes modelos y componentes de enseñanza denominados *Sincrónico* donde el estudiante a través de la tecnología puede unirse a las clases virtuales en tiempo real desde cualquier parte del mundo; *Asincrónico*, cuando el estudiante a través de una plataforma LMS (Sistema de Gestión de Aprendizaje), puede estudiar a cualquier hora con el material disponible en el sistema; y el último componente de aprendizaje es el *Presencial*, que permite que el alumno vaya a una interacción directa con el profesor.

Con el avance de la tecnología, cada momento se genera volúmenes de datos de forma digital, y que mejor, utilizar la técnicas y algoritmos aprendidos en esta maestría de ciencia de datos, para aplicar en las diferentes etapas que con lleva la planificación académica.

En el proceso de este trabajo se realizará entrevistas a los encargados de la planificación institucional y académica, con el fin de entender los pasos y metodologías que utilizan en la construcción de esta tarea que deben hacerlo a inicios de cada periodo académico, se modelara las tablas necesarias para que la información pueda ser tabulada e importada a las herramientas de Python.

Una vez que se termine de modelar y entender, se llevara a Jupyter Notebook el conjunto de datos, para a través de modelos y técnicas de minería de datos pronosticar la oferta distributivo y horario académico.

ABSTRACT

After the pandemic, higher education institutions supported by the control bodies of Ecuador have adopted different models and teaching components called *Synchronous* where the student through technology can join virtual classes in real time from anywhere in the world; *Asynchronous*, when the student through an LMS platform (Learning Management System), can study at any time with the material available in the system; and the last learning component is the *Face-to-face*, which allows the student to have a direct interaction with the teacher.

With the advancement of technology, every moment volumes of data are generated digitally, and what better way to use the techniques and algorithms learned in this master's degree in data science, to apply in the different stages of academic planning.

In the process of this work, interviews will be carried out with those in charge of institutional and academic planning, to understand the steps and methodologies used in the construction of this task, which must be done at the beginning of each academic period, the tables will be modeled. necessary so that the information can be tabulated and imported into Python tools.

Once the modeling and understanding is finished, the data set will be taken to Jupyter Notebook, to forecast the distributive offer and academic schedule through models and data mining techniques.

1. Capitulo I. Introducción

1.1. Generalidades

1.2. Planteamiento del problema

Las IES ofrecen este momento carreras de Pregrado y Postgrado, mismas que por el tema de la pandemia necesitan estructurar y ofertar un modelo de oferta académica de acuerdo con las nuevas normativas de educación superior, esto hace que necesite un sistema y modelo de datos que le permita realizar la Oferta Académica, Distributivo Académico y Horario Académico.

Dentro de esto también se encuentran los componentes de aprendizaje, sincrónico, asincrónico y presencial, donde se debe garantizar la normativa del reglamento de régimen académico vigente en el Ecuador.

1.3. Objetivos

1.3.1. Objetivo General

Realizar una predicción y algoritmo que permita mejorar la oferta, distributivo, y horario de la carrera de enfermería de una institución de educación superior.

1.3.2. Objetivos Específicos

- Realizar un análisis de la data para determinar la cantidad de estudiantes y materias que necesitan proyectarse en cada periodo académico.
- Construir un modelo de datos que permita predecir la oferta académica, el distributivo académico y el horario académico.
- Evaluar los resultados que generen los horarios de la carrera en estudio, diferenciando los componentes: sincrónico, asincrónico, y presencial.

1.4. Alcance

La planificación académica es muy importante a la hora de ofertar materias, realizar un distributivo académico y como resultado final generar los horarios, pero no sirve de nada si solo está desarrollado de forma manual y siguiendo un orden, u ofertando todas las materias de la malla académica, por lo tanto, la utilidad de la ciencia de datos radica en su capacidad para extraer información de conjuntos de datos, como es el caso de la creación de modelos predictivos. Según lo antes expuesto, la razón de este proyecto consiste en sugerir la automatización de la planificación de cursos académicos, basándose en una estimación de

la cantidad de estudiantes y las materias que podrían ser cursadas en un período determinado, una vez obtenida está oferta, pueden ser capaces de armar los paralelos o grupos de forma dinámica, para luego de acuerdo al número de paralelos generar un distributivo académico, en base a la experiencia y al historial académico de los docentes a tiempo completo (TC), medio tiempo (MT), y tiempo parcial (TP), una vez concluido estas 2 etapas generar los horarios académicos.

2. Capitulo II. Revisión Literaria

2.1. Fundamentos teóricos

En esta sección se introduce el conjunto de conceptos teóricos necesarios para respaldar el trabajo actual, donde se presenta un análisis del reglamento de régimen académico vigente en Ecuador, mismo que se utiliza para conocer las orientaciones y regulaciones de las funciones sustantivas de la educación superior, todo lo referente a su gestión, en el marco de la normativa del Consejo de Educación Superior (CES).

Según este Reglamento de Régimen Académico en el TITULO III, que se refiere a la DOCENCIA, permite establecer un marco claro y coherente para garantizar la calidad y la equidad en la educación y la formación académica (CES, 2019).

> **"Artículo 26.- Actividades de aprendizaje**. - Las actividades de aprendizaje procuran el logro de los objetivos de la carrera o programa académico, desarrollan los contenidos de aprendizaje en relación con los objetivos, nivel de formación, perfil profesional y especificidad del campo del conocimiento" (CES, 2019).
>
> La organización del aprendizaje, a través de créditos, se podrá planificar en los siguientes componentes:
>
> a) Aprendizaje en contacto con el docente
> b) Aprendizaje autónomo
> c) Aprendizaje práctico - experimental (que podrá ser o no en contacto con el docente, a excepción del campo de la salud que deberá contar con un docente tutor)
>
> **"Artículo 27.- Aprendizaje en contacto con el docente. -** El aprendizaje en contacto con el docente comprende el conjunto de actividades individuales o grupales desarrolladas con intervención o supervisión directa del docente (de forma presencial o virtual, sincrónica o asincrónica) que comprende las clases, tutorías, conferencias, seminarios, talleres, proyectos en aula (presencial o virtual), entre otras, que establezca la IES en correspondencia con su modelo educativo institucional" (CES, 2019).

Las instituciones de educación superior tienen la capacidad de organizar la enseñanza que implica interacción directa con el profesor, lo cual puede llevarse a cabo a través de la tutoría, salvo en el ámbito de la salud. Cada institución de educación superior establecerá

los métodos y requisitos para llevar a cabo la tutoría, con el objetivo de garantizar el cumplimiento de sus objetivos.

> **"Artículo 28.- Aprendizaje autónomo. -** El aprendizaje autónomo es el conjunto de actividades de aprendizaje individuales o grupales desarrolladas de forma independiente por el estudiante sin contacto con el personal académico o el personal de apoyo académico. Las actividades planificadas y/o guiadas por el docente se desarrollan en función de su capacidad de iniciativa y de planificación; de manejo crítico de fuentes y contenidos de información; planteamiento y resolución de problemas; la motivación y la curiosidad para conocer, investigar e innovar; la transferencia y contextualización de conocimientos; la reflexión crítica y autoevaluación del propio trabajo, entre las principales" (CES, 2019).

> **"Artículo 29.- Aprendizaje práctico - experimental**. - El aprendizaje práctico-experimental es el conjunto de actividades (individuales o grupales) de aplicación de contenidos conceptuales, procedimentales, técnicos, entre otros, a la resolución de problemas prácticos, comprobación, experimentación, contrastación, replicación y demás que defina la IES" (CES, 2019).

2.2. Marco Conceptual

En este apartado se exponen las teorías aplicadas y las definiciones empleadas para la determinación del algoritmo de ciencia de datos para una planificación académica automatizada, se describen los términos más significativos para la comprensión del tema de investigación, se identifican los métodos, las técnicas, elementos, modelado de la minería de datos, asimismo el aprendizaje utilizado para la enseñanza de la minería de datos según la oferta académica requerida.

2.2.1. Oferta Académica

Una oferta académica en educación superior se refiere a los programas y cursos que una institución educativa superior ofrece a sus estudiantes para su formación y educación. Esto incluye programas de pregrado, posgrado, especializaciones, diplomados, entre otros, puede variar según la institución y el país, pero suele estar diseñada para cubrir las necesidades de formación de los estudiantes y las demandas del mercado laboral. Por lo tanto, las instituciones pueden ofrecer diferentes programas académicos en diversas áreas de

estudio, como ciencias sociales, ciencias naturales, ingeniería, humanidades, artes, entre otras (Fallarino et al., 2020).

La oferta académica también puede incluir información sobre las condiciones de admisión, los requisitos de graduación, los recursos disponibles para los estudiantes, la duración de los programas, los costos de matrícula y otros aspectos relevantes para la educación superior (Fallarino et al., 2020) .

2.2.2. Distributivo Académico

Un distributivo académico es un documento o plan que detalla los cursos o materias que un estudiante debe tomar para cumplir con los requisitos de graduación de una institución académica (Tingo et al., 2018).

Por lo general, un distributivo académico incluye información sobre los créditos necesarios para graduarse, los requisitos de cursos básicos y electivos, y cualquier otra información relevante para guiar al estudiante a través de su programa de estudios (Tingo et al., 2018).

También consiste en asignar un docente a las materias proyectadas y ofertadas, siguiendo el Reglamento de Carrera y Escalafón del Profesor de Educación Superior, donde se debe tener en consideración los lineamientos citados en dicha normativa en el capítulo 2, artículos del 5 al 13 (Tingo et al., 2018).

2.2.3. Definición de Horario Académico

El Horario Académico es un documento que establece la distribución del tiempo y la organización de las actividades académicas, incluyendo las clases, seminarios, talleres y otros eventos relacionados con la formación académica de los estudiantes. Este documento es elaborado por la institución académica y se basa en el plan de estudios y el distributivo académico (Guerrero, 2018).

El Horario Académico se utiliza para programar las actividades de los profesores y los estudiantes, y debe estar disponible para ellos con anticipación para que puedan planificar sus actividades personales y académicas. En general, el horario académico se establece para un período académico específico, como un semestre o un año escolar, y se actualiza regularmente en función de los cambios en el plan de estudios, el distributivo académico y las necesidades de los estudiantes y profesores (Guerrero, 2018).

La elaboración del Horario Académico es un proceso complejo que requiere la coordinación de varios departamentos y profesionales, como el departamento de registro académico, el departamento de recursos humanos y el cuerpo docente. Para garantizar la eficacia y la eficiencia en el uso del tiempo y los recursos, el horario académico debe ser diseñado cuidadosamente, teniendo en cuenta la disponibilidad y las necesidades de los estudiantes y profesores, la capacidad de las instalaciones, y la carga de trabajo del personal académico (Guerrero, 2018) .

2.2.4. Técnicas de aprendizaje automático y minería de datos

Esta es una técnica dentro del análisis de datos que posibilita identificar conexiones, categorizar y reunir información con el propósito de generar diversos modelos de conocimiento y se dividen en las tres mencionadas a continuación:

Asociación: Esta modalidad de la minería de datos establece vínculos entre un elemento contenido en un conjunto específico de datos. En el ámbito de la minería de datos, se emplean las reglas de asociación con el propósito de examinar y anticipar información (Schab et al., 2019).

Agrupamiento: Esta metodología efectúa la partición de una extensa colección de datos en agrupaciones coherentes y relevantes. El método más comúnmente empleado en minería de datos es la agrupación, dado que a través de esta se tienen oportunidades para llevar a cabo variados análisis predictivos y facilitar la toma de decisiones (Schab et al., 2019).

Clasificación: constituye un género de técnica en la minería de datos que posibilita anticipar el comportamiento futuro de los datos al categorizar los datos preestablecidos. Diversos algoritmos de clasificación de minería de datos son notables, como el Naive Bayes, Support Vector Machine y el Árbol de Decisión. En la Tabla 1. se realiza una comparativa de los diferentes tipos de técnicas de clasificación:

Dentro de esta investigación, se emplea el Support Vector Machine (SVM) para generar modelos predictivos. Este algoritmo de aprendizaje automático con supervisión busca identificar un hiperplano para clasificar conjuntos de datos. Dos categorías de clasificadores SVM son reconocibles (Schab et al., 2019):

- Clasificador Lineal SVM
- Clasificador no lineal SVM.

Tabla 1. Técnicas de clasificación

Algoritmo	Ventajas	Desventajas
Naive Bayes: Un algoritmo basado en probabilidades que estima la ocurrencia y las relaciones entre valores en un conjunto de datos específico.	Simple, muy rápido, predice el resultado correctamente y de buen rendimiento	Necesita una gran cantidad de datos para generar resultados satisfactorios, ya que opera en base a ejemplos al almacenar todas las muestras de entrenamiento.
Support vector machine (SVM): Un algoritmo sólido que mejora significativamente la capacidad predictiva.	Predicción mucho mejor, estimación rápida del objetivo y utiliza menos parámetros.	Costoso Computacionalmente.
Decisión Tree: Un algoritmo que emplea estructuras arbóreas para la toma de decisiones, facilitando la búsqueda del camino óptimo para la clasificación y logrando así obtener resultados más efectivos.	Mayor precisión en los resultados, menor carga computacional, tiempos de compilación del modelo reducidos y mejora en los tiempos de búsqueda.	Sobreajuste, ramas sin contenido y sin relevancia.

2.2.5. Arboles de Decisión

Un árbol de elección representa un algoritmo de aprendizaje supervisado sin parámetros definidos, empleado en la clasificación y regresión. Posee una organización en forma de árbol con niveles jerárquicos, compuesta por un nodo inicial, ramas, nodos intermedios y nodos finales (Ramos, 2019).

La técnica de aprendizaje en un árbol de decisión adopta una aproximación de dividir y conquistar, mediante una búsqueda ávida para identificar los puntos de separación óptimos en la estructura arbórea. Este procedimiento de división se repite de manera recursiva de nivel en nivel, hasta que la mayoría o la totalidad de los registros se encuentren etiquetados bajo categorías específicas. La determinación de si los datos quedan agrupados de forma homogénea o no, depende de la complejidad del árbol de decisiones.

Los árboles de decisión más sencillos en sus ramificaciones son más fáciles de alcanzar puntos de datos en una sola clase. No obstante, a medida que un árbol crece, se complicado mantener esta pureza, lo cual da como resultado que caigan pocos datos en un subárbol determinado. Este fenómeno es conocido como fragmentación de datos y puede llevar en muchas ocasiones a un sobreajuste del modelo.

En consecuencia, los árboles de decisión tienden a favorecer la creación de árboles de tamaño reducido, en línea con el principio de parsimonia que defiende que "las entidades no deben multiplicarse más allá de lo necesario". En otras palabras, los árboles de decisión optan por incorporar complejidad únicamente cuando es esencial, ya que a menudo la explicación más sencilla resulta ser la más efectiva. Para atenuar la complejidad y prevenir el sobreajuste, se recurre comúnmente a la poda, un proceso que elimina las ramas basadas en características de poca relevancia. La evaluación del ajuste del modelo puede llevarse a cabo a través de la validación cruzada. Una alternativa para preservar la precisión de los árboles de decisión es generar un conjunto mediante la aplicación de un algoritmo de bosque aleatorio; esta técnica de clasificación produce predicciones más acertadas, especialmente cuando los árboles individuales carecen de correlación entre sí. (Ramos, 2019).

Aunque los árboles de decisión tienen aplicaciones versátiles, en muchos casos, otros algoritmos tienden a ofrecer un rendimiento superior al de los árboles de decisión. No obstante, destaca la utilidad especializada de los árboles de decisión en la minería de datos y en actividades de exploración de conocimiento (Ramos, 2019).

2.2.6. Modelado de datos

El modelado de datos es el proceso de crear una representación visual y estructurada de una base de datos, que describe la organización, relación y restricciones de los datos almacenados en la misma. El objetivo del modelado de datos es diseñar una base de datos que sea eficiente, precisa, coherente y fácil de mantener (Ramos, 2019).

En el modelado de datos, se utilizan diferentes tipos de modelos, como el modelo conceptual, el modelo lógico y el modelo físico, para representar diferentes aspectos de la base de datos. El modelo conceptual describe la estructura de alto nivel de la base de datos y se enfoca en los conceptos y relaciones generales. El modelo lógico se enfoca en la estructura detallada de la base de datos y define las tablas, relaciones, restricciones y atributos. El modelo físico se enfoca en cómo se implementa la base de datos en un sistema

de gestión de bases de datos específico, y define la estructura de almacenamiento físico de los datos (Ramos, 2019).

El modelado de datos es una actividad fundamental en el proceso de desarrollo de software y sistemas de información, y se utiliza en diferentes etapas del ciclo de vida del software, desde la planificación y análisis de requisitos hasta el diseño y la implementación (Ramos, 2019) .

2.3. Estructura de Oferta Académica

La oferta académica incluye información detallada sobre los cursos disponibles, sus descripciones, requisitos, horarios, modalidades de estudio (presencial, en línea o semipresencial), créditos académicos, fechas de inicio y duración, entre otros aspectos relevantes. También puede contener información sobre los profesores o instructores, los objetivos educativos y los recursos disponibles para apoyar el aprendizaje.

La estructura de la oferta académica es fundamental para que los estudiantes y participantes puedan tomar decisiones informadas sobre sus estudios, seleccionar los cursos que mejor se adapten a sus necesidades e intereses, y planificar su trayectoria educativa de manera efectiva. Asimismo, es una herramienta esencial para la gestión y administración de los programas educativos por parte de las instituciones.

2.3.1. Criterios para la generación de la oferta académica.

La formulación de una oferta educativa se fundamenta en la pertinencia, la cual está intrínsecamente ligada a la calidad. Una carrera o programa educativo considerado pertinente guarda una estrecha conexión con el entorno en el que opera, ya sea a nivel local, regional o nacional (Paredes et al., 2020).

2.3.2. Técnicas para generar una oferta académica

El proceso de construcción de la oferta académica de una IES (Institución Educación Superior), se puede decir que se desarrolla en dos fases. Se tiene que identificar y erigir los requisitos, con la finalidad de articular la oferta con las necesidades reales de los estudiantes, los requerimientos institucionales y la oferta académica de la población estudiantil nueva (Paredes et al., 2020).

La segunda fase consiste en la construcción del distributivo y horario académico bajo el cumpliendo de los lineamientos del CES, mediante la comisión o departamento de planificación de la IES (Paredes et al., 2020).

Se llevó a cabo una investigación descriptiva de tipo no experimental, con un enfoque cualitativo y con el propósito de aplicar los resultados. Dado que no se manipularon intencionalmente las variables, se obtuvo información de los departamentos y carreras de manera imparcial y objetiva (Paredes et al., 2020).

Se hizo un análisis documental para identificar las normativas, resoluciones y visión de la institución. Se aplicó una encuesta a los estudiantes de 2 carreras con mayor población estudiantil se aplicó́ para identificar su preferencia por los horarios y jornadas de su preferencia (Paredes et al., 2020).

2.4. Métodos de Optimización

En disciplinas como las ciencias empíricas, ciencias de la computación, matemáticas, estadísticas y economía, entre otras, se emplea la optimización matemática o programación matemática para seleccionar el elemento óptimo (según ciertos criterios) de un conjunto dado de elementos disponibles. La optimización está en la base de la investigación operativa, un campo de las matemáticas.

En su forma más básica, un problema de optimización busca maximizar o minimizar una función real al escoger valores de entrada de manera sistemática, dentro de un conjunto permitido, y luego calcular el valor de la función. La teoría de la optimización se generaliza para abordar otras formulaciones, abarcando una amplia área de las matemáticas aplicadas. En términos generales, la optimización implica encontrar los "mejores valores" de una función objetivo en un dominio predefinido, abarcando diversos tipos de funciones objetivo y dominios.

El concepto de optimización se refiere a la acción y el resultado de optimizar. En un sentido más amplio, se relaciona con la habilidad de llevar a cabo o resolver una tarea de la manera más eficiente posible y, en la medida de lo posible, utilizando la mínima cantidad de recursos disponibles.

Durante las últimas décadas, el término "optimización" se ha asociado principalmente con el campo de la informática, pero en realidad, es un concepto que tiene aplicaciones en diversas áreas como las matemáticas, la gestión de procesos y la economía.

2.4.1. Librerías de Python para problemas de optimización

Dentro del sistema científico de Python vamos a encontrar varias librerías que nos van ayudar a enfrentar los problemas de optimización, la que se utilizó en este proyecto es:

Pyomo: Esta librería también nos va a proporcionar un lenguaje para modelar problemas de optimización en Python, este paquete de librería requiere la instalación de diferentes solvers para poder resolver los problemas de optimización, en este caso se utilizó CBC (Pyomo, 2021).

CBC: El solucionador Coin Branch and Cut (CBC) es un solucionador de programa de enteros mixtos (MIP) de código abierto escrito en C++. CBC está diseñado para usarse principalmente como una biblioteca a la que se puede llamar para crear solucionadores personalizados de ramificación y corte. También está disponible una versión ejecutable básica e independiente. CBC es un proyecto activo de código abierto dirigido por John Forrest (coin-or.org, 2005).

2.5. Metodología de Minería de Datos

Oviedo y Jiménez (2019) establecen y detallan cuatro categorías de análisis de datos, que se explican en los siguientes términos:

1. **Analítica Descriptiva**: Dentro de este enfoque de análisis de datos, se suministra información acerca del historial de los datos, los cuales señalan si algo está en un estado positivo o negativo, sin profundizar en las causas subyacentes. Este tipo de análisis se emplea como indicadores que ofrecen una visión retrospectiva.
2. **Analítica Diagnostica:** Esta modalidad de análisis de datos es aún más exhaustiva que la previa, ya que los datos históricos se contrastan con otros datos con el fin de descifrar las razones detrás de los eventos. Mediante este enfoque analítico, es posible examinar interdependencias y tendencias dentro de los datos.
3. **Analítica Predictiva:** En este enfoque analítico, de naturaleza probabilística, se evalúa la probabilidad de que ocurra un evento específico. Se basa en las dos categorías previas de análisis (descriptivo y diagnóstico) para identificar tendencias en los datos, agrupaciones y excepciones. Este tipo de análisis de datos se convierte en una herramienta sumamente valiosa para anticipar eventos o situaciones. No obstante, es crucial resaltar que, a pesar de su relevancia y utilidad para las empresas, esta forma de análisis se trata de una estimación condicionada por el tratamiento y la calidad de los datos.

4. **Analítica Prescriptiva:** Este tipo de analítica permite prescribir las acciones a tomar para problemas futuros o aprovechar las tendencias generadas por la analítica predictiva, este análisis ayuda a la toma de decisiones de las empresas u organismos para mejorar sustancialmente mejores tendencias y minimizar los riegos de los problemas que presente la empresa. La analítica Prescriptiva requiere además de los datos históricos, información externa debido a la naturaleza de los algoritmos estadísticos. Las herramientas y técnicas que utiliza este tipo de analítica son, por ejemplo: machine learning, reglas de negocio y algoritmos. Este tipo es el más sofisticado, lo que hace que sea robusto, pero también más complicado de realizar por las empresas ya que requiere un mayor esfuerzo y más recursos para lograr un valor agregado.

   Dentro del ámbito de la Analítica Prescriptiva, se encuentra una categoría más amplia denominada **Inteligencia Artificial** (IA), que se refiere a máquinas capaces de exhibir comportamientos inteligentes. Una de las manifestaciones de la IA es el Aprendizaje Automático **(Machine Learning**), que forma parte de la inteligencia artificial y tiene la capacidad de aprender de manera autónoma. En esta misma línea, se ubica el Aprendizaje Profundo (**Deep Learning**), una subdivisión del aprendizaje automático que hace uso de redes neuronales (Oviedo & Jiménez, 2019).

   Aparte de estas modalidades de análisis de datos, existen diversas metodologías para adquirir conocimiento, como el proceso KDD (Knowledge Discovery in Databases) y MIDANO. KDD representa un enfoque que establece un procedimiento complejo para identificar patrones nuevos, valiosos y auténticos, con el fin de lograr una comprensión más profunda de los datos (Oviedo & Jiménez, 2019).

3. Capítulo III. Metodología de la Investigación.

3.1. Metodología

El estudio actual adopta un enfoque metodológico combinado que abarca tanto aspectos cualitativos como cuantitativos. Esto se debe a que el análisis del tema de investigación, que en este caso es la oferta académica, se realiza mediante el juicio objetivo del investigador. En relación con la dimensión cuantitativa, se presentarán las matrices y los resultados derivados del análisis de datos.

3.1.1. Método

Se utiliza la metodología CRISP-DM, principalmente debido a su capacidad de distinguir entre los objetivos del negocio y de la minería de datos. Esta metodología aborda todo el flujo de trabajo dentro del ejercicio de la minería de datos, por lo que se utilizará durante todo el trabajo de memoria de título, desde la obtención y recolección de datos, hasta la fase de modelamiento

Este modelo define un proceso de seis fases para la gestión y el análisis de proyectos de minería de datos. Las seis fases son las siguientes:

1. Comprensión del problema: Esta fase implica definir los objetivos del proyecto, determinar los recursos necesarios y establecer un plan de proyecto.
2. Comprensión de los datos: En esta fase se recopilan, seleccionan y describen los datos relevantes para el proyecto.
3. Preparación de los datos: Esta fase implica la limpieza, transformación y manipulación de los datos para asegurar su calidad y su adecuación para el análisis.
4. Modelado: En esta fase se seleccionan y se aplican técnicas de modelado para construir un modelo predictivo o descriptivo.
5. Evaluación: Esta fase implica la evaluación del modelo construido para determinar su validez y utilidad.
6. Despliegue: En esta fase se implementa el modelo en el entorno operativo y se realiza el seguimiento del rendimiento del modelo.

3.1.1.1. Comprensión del negocio

Una de las fases clave de CRISP-DM es la comprensión del negocio, que se enfoca en adquirir un conocimiento profundo de los objetivos y requerimientos del negocio antes de comenzar cualquier análisis de datos (Hernández & Martínes, 2019).

Implica una estrecha colaboración entre el equipo de proyecto y los interesados del negocio. Algunas de las actividades comunes en esta etapa incluyen:

a. Determinar los objetivos del negocio: Es fundamental comprender los problemas o desafíos específicos que el negocio desea resolver a través del análisis de datos. Esto implica identificar las metas, los indicadores clave de rendimiento (KPI) y los resultados esperados (Hernández & Martínes, 2019).
b. Establecer el contexto: Comprender el dominio del negocio en el que se aplicará el análisis de datos. Esto implica conocer la industria, las prácticas comerciales, las regulaciones y cualquier otro factor relevante que pueda influir en el análisis (Hernández & Martínes, 2019).
c. Identificar los requisitos: Recopilar los requisitos específicos de los interesados del negocio. Esto incluye las preguntas clave que se espera que el análisis de datos responda, las necesidades de información, las limitaciones y las restricciones (Hernández & Martínes, 2019).
d. Evaluar la disponibilidad de datos: Analizar la disponibilidad, calidad y relevancia de los datos necesarios para el análisis. Esto puede requerir la colaboración con profesionales de datos para determinar qué datos están disponibles, cómo se almacenan y cómo se pueden acceder (Hernández & Martínes, 2019).
e. Determinar el enfoque de modelado: Basado en la comprensión del negocio y los requisitos identificados, decidir el enfoque de modelado más adecuado. Esto implica seleccionar las técnicas de minería de datos apropiadas y determinar cómo se pueden aplicar para resolver los problemas del negocio (Hernández & Martínes, 2019).

3.1.1.2. Comprensión de los datos

Se centra en obtener una visión profunda de los datos disponibles para el proyecto de minería de datos. Esta fase es crucial para comprender la calidad de los datos, identificar problemas y descubrir patrones relevantes que puedan ser utilizados en el análisis posterior.

Por otro lado, es esencial para establecer una base sólida para el análisis posterior. Proporciona una visión detallada de los datos, su calidad y su relevancia para el proyecto, lo que ayuda a definir estrategias y enfoques adecuados para la siguiente fase del proceso de minería de datos.

3.1.1.3. Preparación de los datos.

Se enfoca en preparar los datos de manera adecuada para el análisis posterior. Esta fase implica realizar tareas de limpieza, transformación y selección de datos con el objetivo de obtener un conjunto de datos limpio, coherente y apto para el análisis de minería de datos (Hernández & Martínes, 2019).

Dentro de las actividades principales que se llevan a cabo durante la fase de Preparación de los datos son (Hernández & Martínes, 2019):

a. Limpieza de datos: En esta etapa, se realizan actividades para identificar y corregir problemas en los datos, como valores faltantes, valores atípicos, errores de entrada o duplicados. Esto implica tomar decisiones sobre cómo manejar los datos faltantes o inconsistentes y eliminar los datos duplicados o incorrectos.
b. Integración de datos: Si existen múltiples fuentes de datos, es posible que se deba integrar la información en un único conjunto de datos coherente. Esto puede requerir la unión de tablas, la fusión de datos de diferentes formatos o la normalización de los valores para garantizar la consistencia.
c. Transformación de datos: Durante esta actividad, se realizan cambios en los datos para adaptarlos a las necesidades específicas del análisis. Esto puede incluir la normalización de variables, la discretización de variables continuas, la creación de nuevas variables derivadas o la agregación de datos a un nivel superior.
d. Selección de atributos: En esta etapa, se decide qué atributos o variables serán incluidos en el análisis final. Esto implica seleccionar los atributos más relevantes para los objetivos del proyecto y descartar aquellos que no aporten valor o estén altamente correlacionados con otros atributos.
e. Formateo de datos: Los datos se formatean de acuerdo con los requisitos del modelo o algoritmo que se utilizará en el análisis posterior. Esto puede implicar la codificación de variables categóricas, la estandarización de las escalas de las variables numéricas o la creación de conjuntos de datos de entrenamiento y prueba.
f. Documentación de los procesos de preparación: Durante todas estas actividades, se documentan los pasos realizados, las decisiones tomadas y cualquier transformación aplicada a los datos. Esto es importante para garantizar la trazabilidad y reproducibilidad del proceso de preparación de los datos.

3.1.1.4. Modelado

Se centra en construir y evaluar modelos de minería de datos utilizando los datos preparados en las etapas anteriores. Esta fase implica la selección de técnicas y algoritmos adecuados, la construcción de modelos, su evaluación y ajuste para obtener resultados óptimos (Pando & Zarate, 2021).

Las actividades principales son (Pando & Zarate, 2021):

a. Selección de técnicas de modelado: En esta etapa, se seleccionan las técnicas y algoritmos de minería de datos más adecuados para abordar los objetivos del proyecto y los requisitos del negocio. Esto puede incluir algoritmos de clasificación, regresión, agrupamiento, asociación u otros, según la naturaleza del problema y los datos disponibles.
b. Diseño y construcción de modelos: Se construyen los modelos utilizando las técnicas seleccionadas. Esto implica aplicar los algoritmos elegidos a los datos preparados y ajustar los parámetros según sea necesario. Durante este proceso, se pueden explorar diferentes configuraciones y enfoques para obtener los mejores resultados.
c. Evaluación de modelos: Se evalúan los modelos construidos para determinar su rendimiento y calidad. Esto implica utilizar medidas de evaluación específicas según el tipo de problema, como precisión, exactitud, sensibilidad, especificidad o índices de error. La evaluación ayuda a identificar los modelos más efectivos y a descartar aquellos que no cumplen con los criterios de calidad establecidos.
d. Ajuste de modelos: Si los modelos no cumplen con los requisitos o no ofrecen un rendimiento satisfactorio, se realizan ajustes para mejorar su desempeño. Esto puede incluir la selección de variables más relevantes, la modificación de los parámetros del algoritmo o la exploración de diferentes enfoques de modelado.
e. Validación cruzada y pruebas adicionales: Se realiza la validación cruzada de los modelos seleccionados para verificar su robustez y generalización. Además, se pueden realizar pruebas adicionales utilizando conjuntos de datos independientes o utilizando técnicas de validación externa para garantizar la confiabilidad de los modelos.
f. Documentación de los resultados y conclusiones: Durante todas estas actividades, se documentan los resultados obtenidos, las decisiones tomadas, las métricas de evaluación y cualquier ajuste realizado en los modelos. Esto es fundamental para comunicar los hallazgos y las conclusiones del proceso de modelado a los interesados y para permitir la reproducción del trabajo en el futuro.
g. La fase de Modelado en CRISP-DM es esencial para construir modelos predictivos o descriptivos basados en los datos disponibles. Permite aprovechar las técnicas y algoritmos de minería de datos adecuados para abordar los objetivos del proyecto

y proporciona una base sólida para las etapas posteriores de evaluación y despliegue del modelo.

3.1.1.5. Evaluación

Tiene como objetivo evaluar de manera exhaustiva los modelos construidos durante la fase de modelado. En esta fase se determina si los modelos cumplen con los objetivos del proyecto y los requisitos del negocio, y si son lo suficientemente precisos y confiables para su implementación.

Las actividades principales que se llevan a cabo durante la fase de Evaluación en CRISP-DM:

a. Evaluación del rendimiento del modelo: Se realiza una evaluación detallada del rendimiento de los modelos construidos utilizando métricas y medidas de evaluación apropiadas para el tipo de problema abordado. Esto puede incluir la precisión, la exactitud, la sensibilidad, la especificidad, el F1-score, el área bajo la curva ROC (AUC-ROC) u otras medidas relevantes (Pando & Zarate, 2021).
b. Comparación de modelos: Si se han construido múltiples modelos durante la fase de modelado, se comparan y contrastan entre sí para identificar el modelo que ofrece el mejor rendimiento en términos de las métricas de evaluación establecidas. Esto ayuda a seleccionar el modelo más adecuado para su implementación (Pando & Zarate, 2021).
c. Evaluación del impacto del modelo: Se evalúa el impacto potencial de implementar el modelo en el negocio. Esto implica analizar cómo el modelo puede contribuir a la solución de los problemas identificados, mejorar la toma de decisiones o generar valor para la organización. También se consideran los riesgos y desafíos asociados con la implementación del modelo (Pando & Zarate, 2021).
d. Validación externa: Se realiza una validación externa del modelo utilizando conjuntos de datos independientes para verificar su rendimiento y generalización en situaciones no vistas anteriormente. Esto ayuda a evaluar la capacidad del modelo para hacer predicciones precisas y confiables en nuevos datos (Pando & Zarate, 2021).
e. Ajuste del modelo: Si el rendimiento del modelo no cumple con los requisitos establecidos, se pueden realizar ajustes adicionales para mejorar su rendimiento. Esto puede implicar la optimización de los parámetros del modelo, la selección de diferentes técnicas de modelado o la incorporación de nuevos datos (Pando & Zarate, 2021).

f. Documentación de los resultados: Durante todas estas actividades, se documentan los resultados de la evaluación, las decisiones tomadas y cualquier ajuste realizado en los modelos. Esto proporciona una referencia clara y completa de los resultados obtenidos y las conclusiones alcanzadas durante la evaluación (Pando & Zarate, 2021).
g. La fase de Evaluación en CRISP-DM es fundamental para garantizar que los modelos construidos sean adecuados y confiables para su implementación en el negocio. Permite tomar decisiones informadas sobre la efectividad y el impacto potencial de los modelos y proporciona una base sólida para la etapa final de implementación o despliegue del modelo (Pando & Zarate, 2021).

3.1.1.6. Despliegue

Se enfoca en implementar y poner en producción los modelos y los resultados obtenidos durante las fases anteriores del proceso. Esta fase implica llevar los resultados del análisis de datos a la práctica y hacer que estén disponibles para su uso en el entorno empresarial (Huancas & Vargas, 2021).

Las actividades que se pueden considerar son (Huancas & Vargas, 2021):

a. Planificación de la implementación: Se elabora un plan detallado para la implementación del modelo, incluyendo los recursos necesarios, los plazos, los responsables y los pasos a seguir. Esto implica definir las acciones requeridas para integrar el modelo en los sistemas existentes o establecer nuevos sistemas.
b. Creación de prototipos: En algunos casos, es útil desarrollar un prototipo o versión inicial del sistema para probar y validar el modelo antes de su implementación completa. Esto puede involucrar la creación de una interfaz de usuario, integración con bases de datos u otros componentes técnicos necesarios.
c. Integración en sistemas empresariales: Se realiza la integración del modelo y los resultados obtenidos en los sistemas empresariales existentes. Esto puede implicar la integración con sistemas de gestión de bases de datos, sistemas de toma de decisiones o cualquier otra infraestructura necesaria para permitir el uso práctico del modelo.
d. Monitoreo y mantenimiento: Una vez implementado, se establece un proceso de monitoreo continuo para evaluar el rendimiento y la efectividad del modelo en un entorno real. Esto implica medir el desempeño del modelo, detectar desviaciones o problemas, y realizar ajustes o actualizaciones según sea necesario.

e. Capacitación y documentación: Se brinda capacitación adecuada a los usuarios y partes interesadas sobre cómo utilizar el modelo implementado y los resultados obtenidos. Además, se documenta de manera completa y clara el proceso de despliegue, los requisitos técnicos y cualquier información relevante para facilitar su mantenimiento y uso futuro.

f. Evaluación continua: Se evalúa regularmente el desempeño y los resultados del modelo implementado para garantizar que siga cumpliendo con los objetivos del negocio y los requisitos establecidos. Si es necesario, se realizan ajustes o mejoras adicionales para mantener la relevancia y la precisión del modelo a lo largo del tiempo.

La fase de Despliegue en CRISP-DM es esencial para convertir los resultados del análisis de datos en acciones y decisiones concretas en el entorno empresarial. Permite aprovechar los beneficios de los modelos y los conocimientos obtenidos y asegura que se utilicen de manera efectiva para impulsar el negocio y generar valor.

3.1.2. Las herramientas a utilizar

- **Oracle**. - Es una plataforma de administración de bases de datos de naturaleza objeto-relacional, esta herramienta nos da muchas ventajas con respecto a otras bases de datos, es escalable y seguro, con capacidad de alto rendimiento, admite SQL como lenguaje de consulta para interactuar con los datos (Salazar et al., 2022).
- **PL/SQL** es el lenguaje de programación que proporciona ORACLE para extender el SQL estándar con otro tipo de instrucciones y elementos, es así como vamos a poder programar, Procedimientos, Funciones, Triggers, Scripts. Versión de B.D. utilizada para el proyecto: Oracle Data base 21c Express Edición for Linux x64 (OL8) (Salazar et al., 2022).
- **Datos. -** Los datos utilizados fueron migrados a las tablas necesarias del gestor de base de datos desde los libros de trabajo proporcionados en Excel. Está información se encuentra en diferentes tablas, que no son más que estructuras de datos en columnas y filas; cada columna es un campo (atributo) y cada fila, un registro (Salazar et al., 2022).
- **Sistema Operativo. -** Para la instalación de la base de datos se utilizó Oracle Linux 8 (OL 8), se escogió está distribución por que una de las ventajas más importante, es que este sistema es autónomo, se aprovisiona, se ajusta y aplica los parches de seguridad por sí mismo, mientras se está ejecutando, sin la interacción del usuario o administrador (Salazar et al., 2022).

- **Jupyter Notebook** utilizando los lenguajes Python y Pandas, y las correspondientes bibliotecas esenciales para la ejecución del proyecto. Pandas una de las principales virtudes que tiene la librería es la carga de datos, pues permite realizar la carga desde distintos orígenes, entre los que acepta de estos encontramos archivos de texto plano como CSV, ficheros en el extendido formato Excel y cargas directas desde bases de datos SQL, entre otros orígenes de datos. Todas estas fuentes de datos contienen la información en formato tabular y pandas permite representar este tipo de datos a la perfección mediante el uso de su estructura principal, el Data Frame, mismo que es la estructura principal de trabajo en Pandas (Salazar et al., 2022).
- **Pyomo** Es un paquete de software de código abierto basado en Python que admite un conjunto diverso de capacidades de optimización para formular, resolver y analizar modelos de optimización. Una capacidad central de Pyomo es el modelado de aplicaciones de optimización estructurada. Pyomo se puede utilizar para definir problemas simbólicos generales, crear instancias de problemas específicos y resolver estas instancias utilizando solucionadores comerciales y de código abierto. Los objetos de modelado de Pyomo están integrados en un lenguaje de programación de alto nivel con todas las funciones que proporciona un amplio conjunto de bibliotecas de apoyo, lo que distingue a Pyomo de otros lenguajes de modelado algebraico como AMPL, AIMMS y GAMS, Pyomo admite una amplia gama de tipos de problemas, que incluyen:
    - Programación lineal
    - Programación cuadrática
    - Programación no lineal
    - Programación lineal entera mixta
    - Programación cuadrática de enteros mixtos
    - Programación no lineal de enteros mixtos
    - Programación estocástica
    - Programación disyuntiva generalizada
    - Ecuaciones algebraicas diferenciales
    - Programación binivel
    - Programas matemáticos con restricciones de equilibrio

Pyomo también es compatible con el análisis iterativo y las capacidades de secuencias de comandos dentro de un lenguaje de programación con todas las funciones. Además, Pyomo también ha demostrado ser un marco efectivo para desarrollar herramientas de análisis y

optimización de alto nivel. Por ejemplo, el paquete PySP proporciona solucionadores genéricos para la programación estocástica. PySP aprovecha el hecho de que los objetos de modelado de Pyomo están integrados en un lenguaje de programación de alto nivel con todas las funciones, lo que permite la paralelización transparente de los subproblemas mediante las bibliotecas de comunicación paralela de Python (Pyomo, 2021)

3.2. Propuesta

Para este trabajo se utilizará la data de 5 periodos anteriores desde el 2019 hasta el 2022. Para obtener los patrones, tendencias y relaciones útiles se manejará el proceso de Minería de Datos, además se pretende aplicar CRISP-DM para generar el conocimiento y se pretende a partir de las variables de entrada obtener el mejor modelo de datos.

3.3. Aplicación de la metodología CRISP-DM

CRISP-DM es una metodología flexible y adaptable que se puede aplicar a cualquier proyecto de minería de datos. Cada fase es esencial para el éxito del proyecto y se pueden volver a visitar y ajustar según sea necesario durante todo el proceso.

3.3.1. Comprensión del negocio

Uno de los problemas que este momento enfrenta las Instituciones de Educación Superior es no poder determinar que materias se pueden ofertar en determinado periodo, debido a las múltiples variables que pueden intervenir como: políticas de número mínimo de alumnos por materia, número de veces que toma un alumno una materia, máximo de créditos u horas permitidos en un determinado periodo.

3.3.1.1. Objetivos del negocio

Las instituciones de educación superior tienen por objetivo cumplir el presupuesto junto al plan operativo anual (POA), para esto deben asegurarse que se cumpla las políticas de abrir una materia con un número mínimo de estudiantes, que permita tener un balance económico de gastos operacionales, departamentos de apoyo y salario de docentes.

3.3.1.2. Criterios de éxito del negocio

- Para un proyecto exitoso se planea utilizar el mejor modelo de entrenamiento.
- Mejorar la tasa de graduación y matriculados en un determinado periodo.

3.3.1.3. Evaluación de la Situación

- La investigación está orientada a la necesidad de las instituciones de educación superior conocer una proyección y oferta académica.

- Los datos son recabados desde el Sistema Académico.
- Recursos de Hardware y Software. – En el desarrollo del proyecto se emplearon herramientas open source de minería de datos, y la responsabilidad de proveer los equipos utilizados recae en el investigador.

Tabla 2. Herramientas tecnológicas

Tipo	Herramienta	Pago	Libre
Software	Python		SI
	Oracle Express	SI	
	Visual Studio Code		
	Equipo Core I7, 12GRam, Disco		
Hardware	1 Tb	SI	
	Memoria Ram 16 G	SI	

Requisitos, supuestos y restricciones.

- Requisitos
  - Obtención de la información necesaria para el proyecto
  - Asesoría de expertos en minería de datos
  - Autorización de desarrollo del proyecto.
- Restricciones
  - Datos limitados.

3.3.1.4. Objetivos de la minería de datos

- Comprender y procesar gran cantidad de datos
- Determinar qué variables tienen un mayor impacto en la capacidad de predecir la oferta académica.
- Crear un modelo de predicción para la oferta académica.
- Realizar la evaluación de los modelos.

3.3.1.5. Producción de un plan de proyecto

Descrito en el cronograma de desarrollo planificado para esta investigación adjunto en anexos.

3.3.2. Comprensión de los datos

Se adquirirán los datos, se organizarán en forma de tablas, se interpretarán y se llevará a cabo un proceso de depuración de los datos recolectados como parte del proyecto.

3.3.2.1. Recolectar datos

Para comprender el proceso que se realiza en una planificación académica, y lograr automatizar dichas tareas se debe entender el origen de los valores, para eso se utilizó encuestas a las personas que llevan este proceso, preguntándoles algunos parámetros que intervienen en:

1. Proyección académica
2. Oferta académica
3. Distributivo académico
4. Horario académico

Para la proyección académica necesitan poder clasificar a las carreras por su tamaño: Grande, Mediana y Pequeña, lo que les permitirá conocer qué población estudiantil tienen en cada una de ellas, con esto lograr tener un enfoque más preciso de las matrículas y estrategias a sugerir para el periodo a iniciar, todo esto en base al número de matrículas obtenidas en uno o dos periodos anteriores. Además de esto se observa que todo el análisis se centra en el récord académico del alumno, analizando materias tomadas y pendientes, para la proyección se basan en las asignaturas que al alumno le falta por seguir, revisando los pre-requisitos de cada una de ellas.

La Oferta Académica se concentra en crear paralelos con el número de alumnos probables que se vayan a matricular en el periodo, respetando los siguientes parámetros:

- Número máximo de créditos u horas por carreras y periodos.
- Mínimo de estudiantes necesarios por paralelo para que se abra un grupo.
- Número de alumnos por tipo de aprobación (Regular, Tutoría, Examen).

La solución tiene que ser capaz de lograr que un alumno tome la mayor cantidad de materias posibles.

Por último el Horario Académico es el resultado final de todo el proceso de planificación académica, donde se crea los horarios por cada materia y carrera respetando las jornadas establecidas por los estudiantes o el Dpto. de Planificación, entre los parámetros que se sigue es: los componentes de aprendizaje de una materia si es sincrónica, presencial

o asincrónica, aquí también se controla la capacidad de aulas, de tal forma que ningún paralelo pueda sobrepasar el aforo de cada espacio físico.

Se despliega la Tabla 3 de periodos académicos.

Tabla 3. Periodos académicos

No	Fecha Inicio	Fecha Fin	Periodo
1	1 de enero de 2020	30 de junio de 2020	52
2	1 de julio de 2020	31 de diciembre de 2020	53
3	1 de enero de 2021	30 de junio de 2021	54
4	1 de julio de 2021	31 de diciembre de 2021	55
5	1 de enero de 2022	30 de junio de 2022	56
6	1 de julio de 2022	31 de diciembre de 2022	57

Se despliega todos los alumnos de la carrera de Enfermería.

Tabla 4. Población estudiantil con materias por tomar.

Carrera	**Enfermería**
Materias X Nivel	**Números De Cupo**
1	**254**
Bioquímica	24
Comunicación oral y escrita	41
Educación para la salud	23
Enfermería básica	22
Estadística	33
Herramientas informáticas	44
Introducción a la investigación y pensamiento critico	25
Psicología general	42
2	**801**
Atención primaria en salud	130
Bioestadística	149
Ética profesional	93
Farmacología básica	61
Filosofía de la enfermería	99
Nutrición	65
Salud y sociedad	94
Socio antropología	110
3	**521**
Bioética	85

	Desarrollo personal	94
	Epidemiologia	37
	Farmacología aplicada a la enfermería	67
	Fisiopatología	59
	Microbiología y parasitología	18
	Morfofisiología	81
	Semiología	80
4		**616**
	Administración de enfermería	125
	Cuidados a la mujer	127
	Enfermería clínica I	22
	Enfermería quirúrgica I	34
	Enfoque de género aplicados a la salud	130
	Metodología de la investigación	109
	Salud mental	69
5		**523**
	Atención de enfermería en neonatología	123
	Enfermería clínica II	105
	Enfermería comunitaria I	50
	Enfermería pediátrica	135
	Enfermería quirúrgica II	110
6		**1074**
	Cuidados integrales al adolescente	147
	Culturas ancestrales en el ecuador	130
	Enfermería comunitaria II	112
	Enfermería materno infantil I	103
	Enfermería psiquiátrica	144
	Gerencia de la salud publica	136
	Salud laboral	156
	Transculturalidad y salud	146
7		**656**
	Enfermería de urgencias	89
	Enfermería en adicciones	101
	Enfermería materno infantil II	83
	Legislación en enfermería y salud	146
	Producción científica en salud	120
	Saberes ancestrales de la salud	117
8		**172**
	Internado rotativo-en pediatría	150
	Internado rotativo-en salud familiar y comunitaria	22
9		**11**
	Internado rotativo-en ginecoobstetricia y neología normal	11
Total General		**4628**

3.3.2.2. Verificar la calidad de los datos

Utilizando la matriz de correlación, analizamos la relación entre las diferentes variables y comprendemos su nivel de asociación.

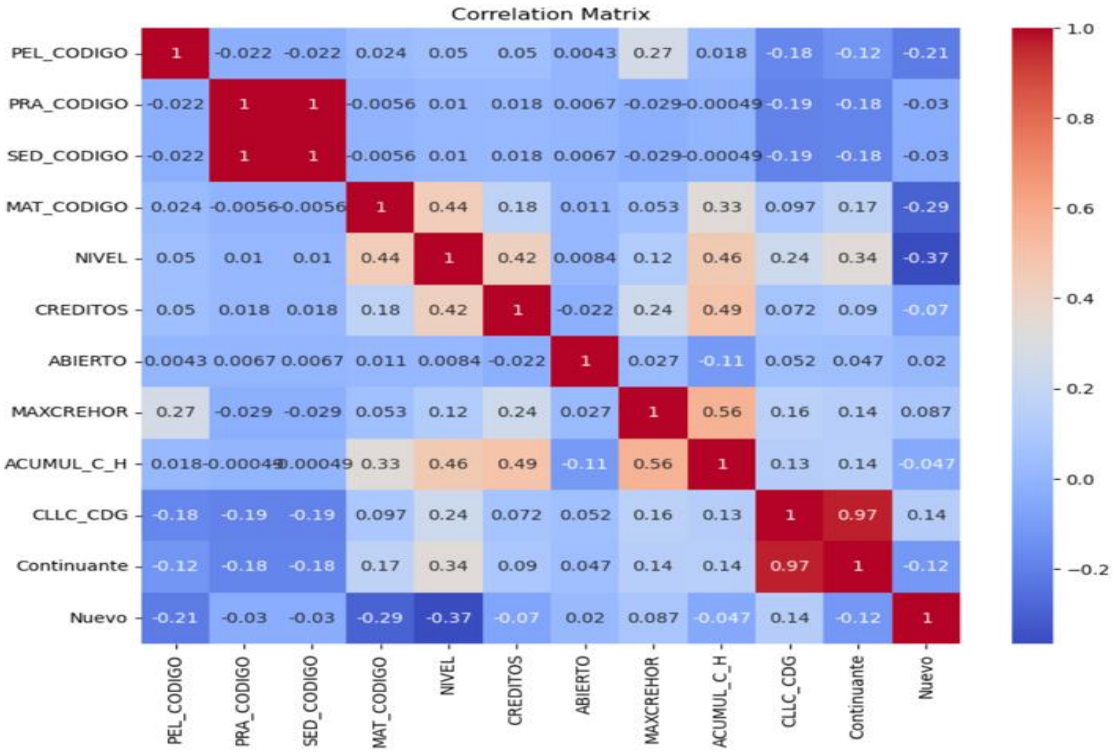

Figura 1. Matriz de Correlación

Del análisis de la matriz de correlación se puede observar que la variable "Abierto" es la que menos asociación tiene, por lo que se procede a eliminar, y además es nuestra columna objetivo.

3.3.3. Preparación de los Datos

Continuando con el objetivo del proyecto: Construir un modelo de datos que permita predecir la oferta académica, el distributivo y el horario académico.

Instalamos las librerías Pyomo y el Solver CBC, para esto seguimos lo comandos tradicionales y comunes.

```
#!pip install pyomo
#!conda install -c conda-forge coincbc
```

3.3.3.1. Selección de datos

Los datos contienen información histórica desde el periodo 54 hasta el 59 de la carrera de salud, para el análisis se va utilizar la data de la carrera de enfermería.

La selección de atributos tiene información relevante y confidencial de los estudiantes durante el proyecto, pudiéndose filtrar las materias, el número de horas, número de créditos máximos por periodo, y si una materia se abrió o no en determinado periodo.

3.3.4. Modelado

Continuando con el objetivo de encontrar el modelo que permita generar la oferta académica, se presenta a continuación el análisis de los modelos a utilizar.

3.3.4.1. Selección de técnicas de modelado

Cuando se trata de seleccionar técnicas de modelado, es importante considerar varios factores, como el tipo de datos, el objetivo del modelado, la disponibilidad de datos y los recursos computacionales disponibles. A continuación, se presentan algunos pasos generales que puedes seguir para seleccionar las técnicas de modelado adecuada.

Para el proyecto se tomó 3 algoritmos para determinar el mejor modelo a seguir en cuanto a la oferta y proyección académica.

- Árboles de decisión
- Support Vector Classifier
- Gaussian Naive Bayes

Con respecto a los horarios se siguió un Modelo de Optimización.

Se consideró el horario académico, en una tabla se asignó una materia para cada día de la semana y hora correspondiente. Además, la variable utilizada para tomar decisiones es de naturaleza binaria, es decir,

$$vbSubjectSchedule[d,h,s]=1$$

sí el día dado “d” y la hora “h”, se enseña la asignatura “s”. En cualquier otro escenario se establecerá como 0.

Lo cual da lugar a 3 conjuntos distintos que se utilizará en el problema:

- $sDays$: días de la semana donde se requiere generar nuevo horario.

- $sHours$: horas del día en la que se imparte asignatura.
- $sSubjects$: asignaturas sé que necesitan asignar.

  Para construir el modelo de Pyomo definimos varios parámetros:
- $max\_hours\_per\_day$: número máximo de horas de la asignatura que se imparte en un día.
- $hours\_per\_subject$: Cantidad de horas por semana asignadas a cada materia.

Posteriormente se establecen los horarios incorporando al modelo diversos parámetros y variables adicionales que ayudarán a modelar restricciones o permitirán acceder de manera eficiente a valores una vez que se haya resuelto el modelo. A partir de este punto en adelante, emplearemos el término "parámetros" para hacer referencia a cualquier elemento que posea un valor predefinido y constante, y utilizaremos el término "variables" para describir los componentes cuyo valor estará sujeto a las decisiones que tomemos.

- $pHoursPerSubjectsSubjects$: cantidad de horas en la semana según la asignatura
- $pMinDaysPerSubjectsSubjects$: número mínimo de días en los que se puede enseñar una asignatura, considerando el total de las horas en la semana y el máximo de horas en el día.
- $pMaxDaysPerSubjectsSubjects$: dependiendo el caso, se limita o no los días y las horas para el año lectivo.

  Variables adicionales o auxiliares.
- $vbSubjectDaysFlagssDays,sSubjects:$ variable binaria utilizada para supervisar si se enseña una materia específica en un día de la semana dado..
- $vbSubjectSwitchessDays,sHours,sSubjects:$ variable binaria indicadora que adoptará el valor 1 si la materia se enseña en la siguiente hora y no en esta, o viceversa, y tomará el valor 0 si ambas asignaciones coinciden.
- $vIsubjectTotalDays:$ variable que suma la cantidad total de días en los cuales se enseña cada materia, en caso de que deseemos priorizar la agrupación de asignaturas en un número reducido de días dentro de las restricciones establecidas.

  Restricciones.

  Se define lo que se va restringir en cuanto a la decisión del modelo matemático.
- C1-No más de 3 materias por hora y no menos de una.

- C2-Número de horas de la asignatura igual al número de horas esperadas.
- C3-Total horas de una asignatura < al máximo de horas de una materia en un día.
- C4-Identificar que días una materia se imparte.
- C5-Verificar que no se dé menos o más días de clases para una materia.

Función Objetivo

Además de su objetivo de cumplir todas las restricciones, el modelo de optimización debe ser capaz de comparar soluciones para identificar la mejor opción. Para lograr esto, evaluaremos la distribución de la semana según la suma de preferencias asignadas (con un valor predeterminado de 1 para todas las asignaciones) y aplicaremos penalizaciones por cada día adicional utilizado en comparación con el mínimo requerido para impartir cada materia.

Formalmente:

$$\max z = \sum_{i \in sDias} . \sum_{j \in sHoras} \sum_{i \in Asignaturas} pPreferencias_{i,j,k} * vbHorarios_{i,j,k} - penalidad * vIDiasTotalAsignaturas$$

4. Capítulo IV. Resultados

4.1. Análisis del estado actual

4.1.1. Comprensión del Negocio

En cada institución existen periodos académicos que se abren de acuerdo con el calendario académico, los mismos que consideran la oferta de asignaturas que sean posible planificar de acuerdo con su duración, con el propósito de mantener la semestralización y las regularizaciones de mallas académicas de los estudiantes, considerando las adecuaciones de los ajustes de ámbito académico realizados en los Planes de Estudio.

En cuanto al distributivo, la cantidad de horas límite de dedicación del estudiante en un periodo académico a las actividades docentes se establecen en la carrera, según lo que establece el RRA/CES. Lo cual es el referente para determinar hasta cuantas horas podrá tomar el estudiante en un periodo académico.

En lo referente a los horarios, se planifican de acuerdo con las horas en las que se produce la interacción en tiempo real (sincrónica en entorno virtual, o presencial en entorno real) entre profesores y estudiantes, lo cual corresponde a las horas de clases en contacto con el docente más las horas de prácticas y experimentación de los aprendizajes que se realizan, de igual manera, de forma sincrónica o de forma presencial.

Las horas en las que no se produce esa interacción sincrónica o interacción presencial en entorno real, es decir, que se realizan de manera asincrónica, no se planifican en el horario.

4.1.2. Problemática a resolver

Para iniciar la investigación se pudo aplicar una entrevista al responsable del área para identificar las necesidades de este y así poder estructurar su situación problemática. Una vez detallado la problemática y propuesto los objetivos, se realizaron las observaciones de campo no experimental para las cuales se utilizaron diarios de trabajo independientes por cada variable de estudio

En el área de planificación se realizó la aplicación del diario de trabajo a la jefa de planificación y calidad, donde se detalla el proceso a seguir. Para el mejor entendimiento de cada proceso se facilitó la información para poder crear un diagrama de flujo del proceso de planificación, según la metodología CRISP-DM. Luego se inició la fase de modelamiento donde se tomó la información ya procesada de los instrumentos y de la base de datos.

En la fase de evaluación se corrieron los modelos predictivos y se hizo la simulación del llenado de los instrumentos en base a los datos arrojados por el modelo, a fin de poder hacer un comparativo de los indicadores del pre-test y del post-test.

CRISP-DM. – Significa en inglés Cross-Industry Standard Process for Data Mining, es un método comprobado que ayuda a orientar los trabajos de minería de datos, también permite distinguir los objetivos del negocio. Está metodología se utilizará durante todo el trabajo, partiendo desde la obtención y recolección de datos, hasta la fase de modelamiento (Salazar et al., 2022).

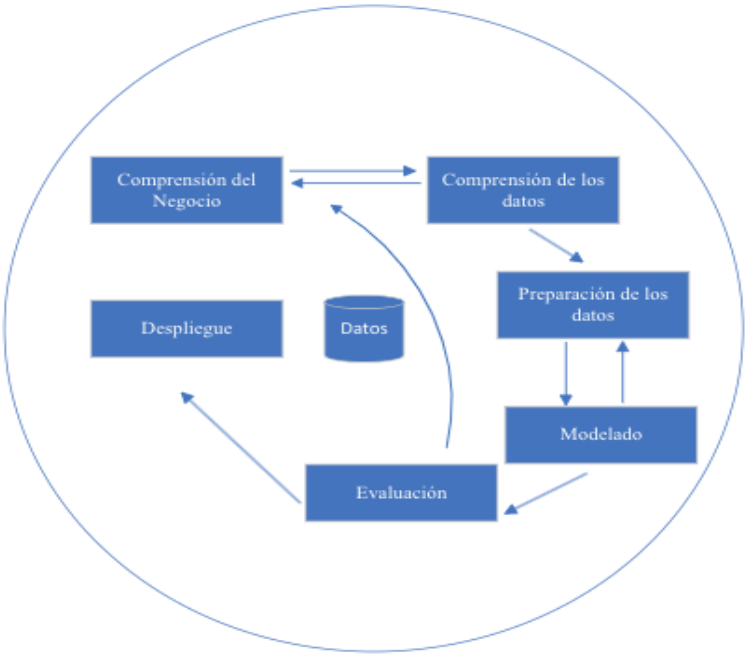

Figura 2. Crisp-DM

4.2. Aplicación de las técnicas de minería de datos.

La planificación académica es el proceso de diseñar y organizar un plan detallado de actividades y metas a seguir en el ámbito académico. Se trata de establecer un conjunto de acciones y objetivos para alcanzar el éxito académico en un período determinado, ya sea a nivel individual, de un departamento o de una institución educativa.

Una planificación académica puede abarcar diferentes aspectos, como el currículo, la secuencia de materias, los recursos necesarios, las estrategias de enseñanza, las evaluaciones, el distributivo y los horarios. Estos elementos se organizan de manera coherente y estructurada para lograr los resultados educativos deseados.

Algunos elementos clave en la planificación académica son:

- Establecimiento de objetivos: Definir claramente los objetivos y metas que se desean alcanzar en el período académico. Estos objetivos pueden estar relacionados con el aprendizaje de los estudiantes, el desarrollo de habilidades, la adquisición de conocimientos específicos o el cumplimiento de requisitos educativos.
- Diseño del currículo: Determinar los contenidos y las habilidades que se deben enseñar en cada materia o curso. Esto implica seleccionar los temas relevantes, organizarlos en una secuencia lógica y establecer los recursos necesarios para su enseñanza.
- Programación de actividades: Establecer un cronograma detallado que indique cuándo y cómo se llevarán a cabo las diferentes actividades académicas, como clases, conferencias, prácticas de laboratorio, exámenes y proyectos. La programación garantiza que los contenidos se aborden en el tiempo asignado y se cumplan los plazos establecidos.
- Selección de métodos de enseñanza: Identificar las estrategias y enfoques pedagógicos más adecuados para transmitir los contenidos y alcanzar los objetivos educativos. Esto puede incluir la selección de materiales didácticos, el uso de tecnología educativa, la implementación de actividades prácticas y el fomento de la participación activa de los estudiantes.
- Evaluación del aprendizaje: Definir los criterios y los métodos de evaluación para medir el progreso y el logro de los estudiantes. Esto implica establecer los tipos de evaluaciones, como exámenes escritos, proyectos, presentaciones o evaluaciones continuas, y determinar cómo se utilizarán los resultados para mejorar el proceso de enseñanza-aprendizaje.
- La planificación académica proporciona una estructura sólida para el desarrollo y la ejecución de actividades educativas. Ayuda a garantizar la coherencia, la eficiencia y la calidad en la enseñanza, y facilita el seguimiento y la evaluación de los resultados educativo.

4.2.1. Comprensión de los datos

Para esto se tomó como ejemplo una planificación de una IES anónima donde se evaluará la oferta, distributivo y horario, también se evaluarán los indicadores que ayudaran a mejorar las proyecciones de la oferta académica, como resultado lo que se busca es obtener los horarios de las asignaturas que probablemente más alumnos se matriculen, esto con el fin de optimizar espacios físicos y claustro docente.

Del análisis realizado se pudo observar que la planificación se divide en 3 procesos:

La oferta académica que permite determinar qué materia se abre en cada periodo académico de acuerdo a la malla y a los pre requisitos académicos, en esta parte es donde se determina cuantos paralelos se puede abrir por cada materia, por otro lado, esto se vincula a la parametrización que se realiza por sede y carrera en cuanto al número de créditos u horas permitidas tomar en cada periodo que se abre.

Tabla 5. Descripción de la tabla oferta.

Campo	**Descripción**
Secuencia	Código de secuencia
cod_materia	Código de materia
cod_nivel	Nivel de materia
cod_sede	Código de la sede
cod_campus	Código campus
cod_facultad	Código facultad
cod_carrera	Código carrera
cod_grupo	Número de grupo o paralelo
descripcion_grupo	Nombre del paralelo o grupo
num_matriculados	Contador de alumnos matriculados
num_cupos	Número de cupos
codigo_periodo	Código de periodo

En el distributivo lo que se busca es asignar a cada paralelo de materias ofertadas un docente, en esta etapa se debe garantizar que se respete el tiempo de dedicación que

tiene estipulado en el contrato, lo que se ha podido también observar, es que por ejemplo siempre falta docentes por la cantidad de materias y paralelos que genera la oferta.

Tabla 6. Descripción de la tabla distributivo.

Campo	Descripción
Secuencia	Código de secuencia
cod_materia	Código de materia
cod_nivel	Nivel de materia
cod_sede	Código de la sede
cod_campus	Código campus
cod_facultad	Código facultad
cod_carrera	Código carrera
cod_grupo	Numero de grupo o paralelo
descripcion_grupo	Nombre del paralelo o grupo
num_matriculados	Contador de alumnos matriculados
codigo_periodo	Código de periodo
cod_docente	Código de docente
cod_contrato	Código de contrato (TC,MT,TP)
tipo_aprendizaje	Código de aprendizaje (S=Sincrónico, A=Asincrónico, P=Presencial)

En el horario académico podemos identificar algunas características importantes que debe manejar en el momento de una planificación académica, además se debe garantizar que los horarios no se crucen para que el estudiante pueda tomar la mayor cantidad de materias, dentro de los campos que se pudo diseñar y modelar están:

Tabla 7. Descripción de la tabla Horario Académico.

Campo	Descripción
Secuencia	Código de secuencia
cod_materia	Código de materia
cod_nivel	Nivel de materia
cod_sede	Código de la sede

cod_campus	Código campus
cod_facultad	Código facultad
cod_carrera	Código carrera
cod_grupo	Numero de grupo o paralelo
descripcion_grupo	Nombre del paralelo o grupo
num_matriculados	Contador de alumnos matriculados
codigo_periodo	Código de periodo
cod_docente	Código de docente
cod_contrato	Código de contrato (TC,MT,TP)
tipo_aprendizaje	Código de aprendizaje (S=Sincrónico, A=Asincrónico, P=Presencial)
hora_inicio	Hora inicial
hora_final	Hora final
cod_aula	Código del aula

Como se puede apreciar las restricciones, llaves primarias y foráneas se mantienen de una tabla a otra, con el fin de garantizar la integridad del modelo.

4.2.2. Recopilación de los datos iniciales

En esta parte se obtuvo los datos iniciales que nos ayudaran a desarrollar nuestro análisis de datos y modelo.

Tabla 8. Descripción de la tabla horario académico

Periodo	Ciudad	Carrera	Número de alumnos
2022	Guayaquil	Enfermería	459

Tabla 9. Número de docentes con tiempo de dedicación

Número docentes	Relación IESS	Categoría	Tiempo de Dedicación
60	Contrato con relación de dependencia	Docente Ocasional	TC
20	Contrato con relación de dependencia	Docente Titular	TC
10	Contrato con relación de dependencia	Docente Titular	MT
5	Contrato sin relación de dependencia	Técnico Docente	TP

Nota: (TC= Tiempo Completo, MT=Medio Tiempo, TP=Tiempo Parcial).

Tabla 10. Espacios físicos en la carrera de Enfermería

Sede	Facultad	Carrera	Nombre Espacio Físico	Código Espacio Físico	Numero Pisos	Capacidad
Quito	Facultad de Salud	Enfermería	Aula 1	A1	1	30
Quito	Facultad de Salud	Enfermería	Aula 2	A1	1	20
Quito	Facultad de Salud	Enfermería	Aula 3	A1	1	15
Quito	Facultad de Salud	Enfermería	Aula 4	A1	1	20
Quito	Facultad de Salud	Enfermería	Aula 5	A1	1	35
Quito	Facultad de Salud	Enfermería	Aula 6	A1	1	20
Quito	Facultad de Salud	Enfermería	Aula 7	A1	1	15
Quito	Facultad de Salud	Enfermería	Aula 8	A1	1	20
Quito	Facultad de Salud	Enfermería	Aula 9	A1	1	30
Quito	Facultad de Salud	Enfermería	Aula 10	A2	2	25
Quito	Facultad de Salud	Enfermería	Aula 11	A2	2	30
Quito	Facultad de Salud	Enfermería	Aula 12	A2	2	20
Quito	Facultad de Salud	Enfermería	Aula 13	A2	2	32
Quito	Facultad de Salud	Enfermería	Aula 14	A2	2	15

4.2.3. Descripción de los datos

Para abordar el trabajo de este capítulo, se obtiene un Excel con todos los campos observados en el entendimiento del negocio, se forma la matriz de la parametrización de la proyección académica.

Tabla 11. Parametrización de datos

Ciudad	Facultad	Carrera	Parámetro	No. Est. Matriculados Periodo Anterior	Ap. Regular	Ap. Tutoría	Ap. Examen
Quito	Facultad de Salud	Enfermería	Carrera Grande	550	15	8	2

A partir del entendimiento generado se obtiene las matrices con los datos de materias y estudiantes en la carrera de estudio.

Como se puede apreciar en los cuadros por cada materia se ha tabulado el número de cupos, que vienen hacer la cantidad de alumnos posibles que tomen la asignatura en el periodo, esto nos permite tener una visión más precisa de nuestra oferta académica.

Teniendo esta información básica que es la parametrización de datos más la cantidad de alumnos por materia, ya podemos construir el algoritmo que nos permita crear los paralelos de forma automática en paralelos Grandes, Medianos y pequeños que permita agrupar las materias con más números de alumnos probables que puedan tomar en un determinado periodo.

La Tabla 12 muestra nombres de variables y descripción de cada una de ellas.

Tabla 12. Descripción de los datos

Campo	Descripción
codigo_periodo	Código de periodo académico, tipo numérico
cod_proyecto	Número de proyecto o malla académica
cod_sede	Código de la sede
cod_campus	Código del campus

cod_facultad	Código de la facultad
cod_carrera	Código de la carrera
nom_carrera	Nombre de la carrera
cod_materia	Código de la materia
cod_alumno	Código del alumno
cod_nivel	Número del nivel de la materia
num_horas	Número de horas
num_veces_tomadas	Número de veces tomada la materia
tipo_unidad	Tipo de crédito u hora
num_horas_max_per	Número máximo de horas que puede tomar un alumno en un periodo académico.
num_horas_max_alumno	Número de horas que va a tomar el alumno en el periodo
alumno_tipo	Tipo de alumno
materia_si_abre_o_no	Código para identificar si se abre o no, donde 1 = si se abre, y nulo = no se abre.

4.2.4. Exploración de los datos

Con el fin de dar cumplimiento al objetivo 2 de "Construir un modelo de datos que permita predecir la oferta académica, el distributivo académico y el horario académico".

Se inicia el proceso de cargar el dataframe utilizando el lenguaje de programación Python, incorporando las bibliotecas requeridas para llevar a cabo un análisis exhaustivo de las variables de mayor relevancia.

Con el comando pd.read_excel cargamos a un *dataframe* nuestros datos en la Figura 3 se despliega los datos leídos desde Python.

	codigo_periodo	cod_proyecto	cod_sede	cod_campus	cod_facultad	cod_carrera	nom_carrera	cod_materia	cod_alumno	cod_nivel	num_horas	num_veces_tomadas
0	54	89	1	1	17	2	EFERM	1636	28689	6	80	0
1	55	89	1	1	17	2	EFERM	1640	28689	7	80	0
2	54	89	1	1	17	2	EFERM	1638	28689	7	160	0
3	54	89	1	1	17	2	EFERM	1642	28689	7	120	0
4	54	89	1	1	17	2	EFERM	1641	28689	7	120	0
...	...	...	...	...	...	...	...	...	...	...	...	...
187	54	90	2	6	17	2	EFERM	1013	42422	1	80	0
188	54	90	2	6	17	2	EFERM	1601	42422	1	120	0
189	54	90	2	6	17	2	EFERM	1018	42422	1	80	0
190	54	90	2	6	17	2	EFERM	1026	42422	1	80	0
491	54	90	2	6	17	2	EFERM	1602	42422	1	200	0

Figura 3. Carga y exploración del dataset

Dentro de la data cargada tenemos una columna “SI SE ABRIO”, como se puede ver esta variable tiene la letra “S” cuando una materia se abre y “N” cuando no se abre (Figura 4).

UNIDAD_MATERIA	MAXCREHOR	ACUMUL_C_H	TIPO_ALUMNO	SI SE ABRIO	ABIERTO
H	850	80	Continuante	S	1
H	300	80	Continuante	S	1
H	850	600	Continuante	S	1
H	850	840	Continuante	S	1

Figura 4. Describe si una materia se abrió "S", caso contrario "NULL"

A continuación, se revisa con el método describe de Python, que devuelve información estadística de los datos del *dataframe*. Esta información incluye el número de muestras, el valor medio, la desviación estándar, el valor mínimo, máximo, la mediana y los valores correspondientes a los percentiles 25% y 75%.

Tabla 13. Valores obtenidos con el comando describe

Estadístico	**Periodo**	**Máximo de número de horas por periodo**	**Número de horas por alumno**	**Estado de materias abiertas**
Count	24492.00	24492.00	24492.00	24492.00
Mean	56.35	810.31	441.13	0.51

Std	1.69	154.39	249.14	0.50
Min	54	300.00	80.00	0.00
25%	55	850.00	240.00	0.00
50%	56	850.00	400.00	1.00
75%	58	850.00	640.00	1.00
Max	59	1560.00	1560.00	1.00

Por otro lado, dentro del DF importado a Python se obtiene valores de tipo string para lo cual se realizar una conversión a valores numéricos, en este caso se aplica a las siguientes variables:

"SI SE ABRIO ", Variable que identifica si una materia se abrió en un determinado periodo, S=1, N=0.

"CONTINUANTE O NUEVO", Variable que identifica si un alumno es nuevo en la carrera o ya viene cursando periodos anteriores, Continuante =1, Nuevo =0, en este caso se divide en 2 columnas.

Con la ayuda de las funciones barplot, boxplot, violinplot, se procede a explorar las variables del proyecto.

De acuerdo a la Figura 5, lo primero que vamos a revisar es la relación de las materias abiertas según los periodos académicos.

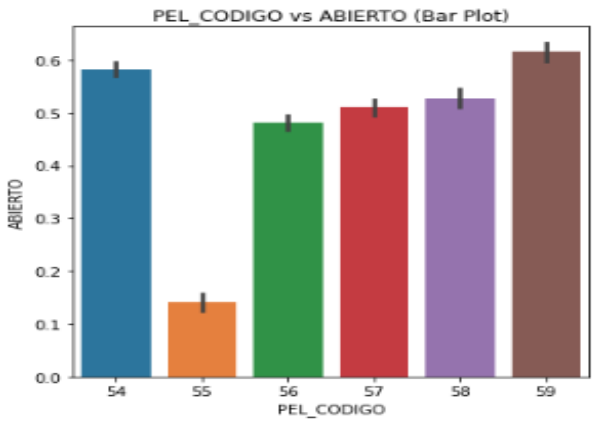

Figura 5. Gráfico de barras, representa número de materias abiertas según los periodos

En la figura 6 se observa que en el periodo 55 existe un *outlier* debido a que hubo un cambio de mallas, por ende, para el análisis no se tomaría encuentra el periodo 54 y 55, se utilizara como referencia el periodo 56,57,58,59.

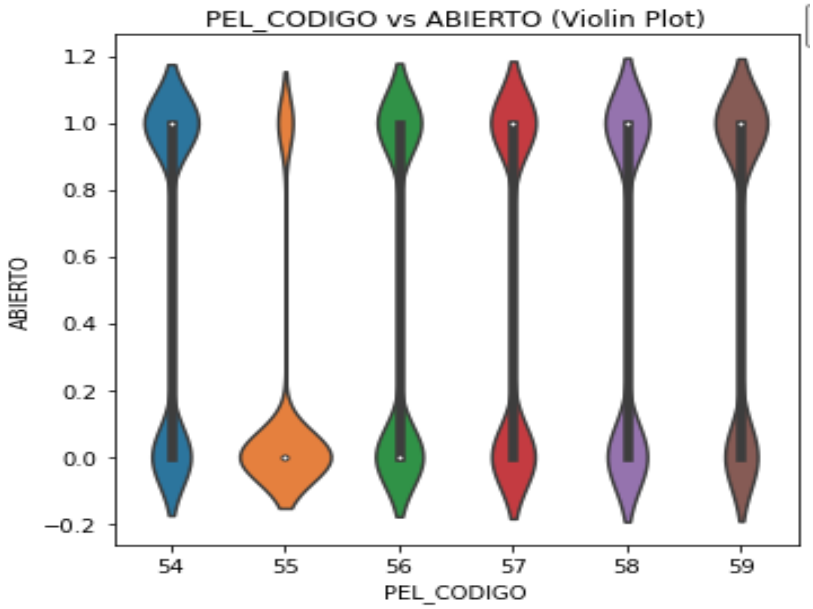

Figura 6. Gráfico violín, indica el número de cursos abiertos según los periodos.

En la figura 7 se puede apreciar las materias que casi siempre se abren, y las que tienen una probabilidad baja de abrirse.

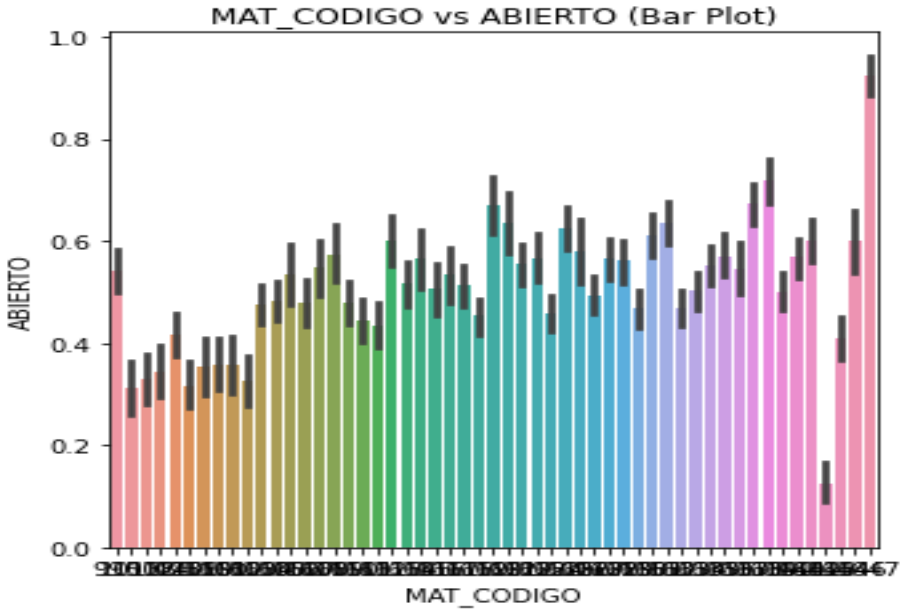

Figura 7. Gráfico de materias que se han abierto

En el siguiente grafico de la Figura 8, se puede observar las materias de que nivel se han abierto con más frecuencia.

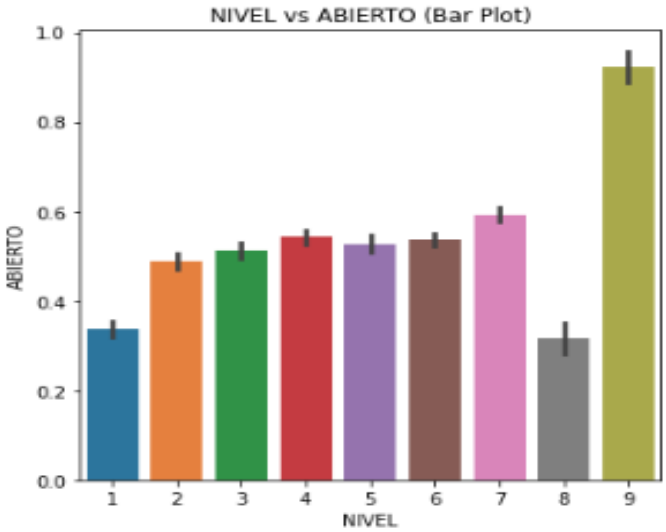

Figura 8. Gráfico de materias por el nivel que se han abierto

En la Figura 9 se observa las materias con más horas que frecuentemente se abren, por ejemplo, las materias con 80 horas tienen una frecuencia de abrirse de 45%, en cambio las materias de 200 horas tienen una frecuencia de abrirse del 58%, teniendo como observación que las materias con 800 horas frecuentemente no se abren ya que tiene una frecuencia 1.10% de abrirse, debido a que son materias de internados rotativos.

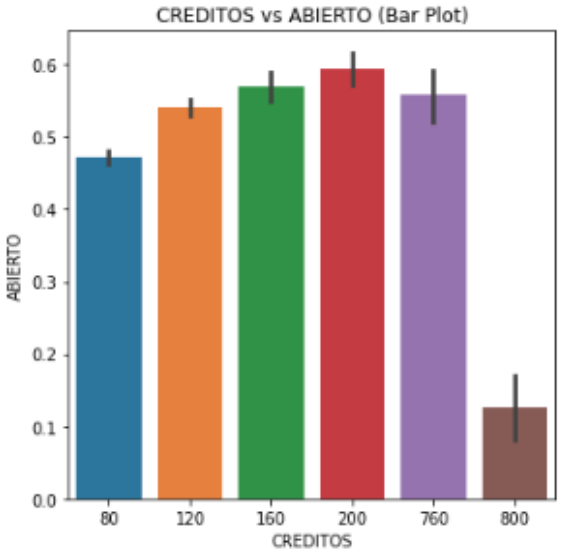

Figura 9. Gráfico de materias por nivel que se han abierto.

A continuación, se utilizó una herramienta importante en el ámbito de la inteligencia artificial y el aprendizaje automático que es la **matriz de confusión** que nos permite calcular los verdaderos positivos (TP) y negativos (TN), y los falsos positivos

(FP) y negativos (FN), para lograr obtener el Accuracy (Exactitud), y la Precisión (Precision).

La Exactitud (Accuracy) nos permitió identificar la cantidad de predicciones positivas que fueron correctas.

La Precisión (Precisión) nos dio el porcentaje de casos positivos detectados.

4.2.5 Verificación de la calidad de los datos

En esta labor, se llevaron a cabo comprobaciones en los datos con el propósito de evaluar la coherencia de los valores individuales en los campos, la frecuencia y distribución de los valores faltantes, así como la identificación de valores que excedieran los límites establecidos. Estos últimos podrían considerarse como datos no deseados que podrían afectar el proceso.

Con el comando df.isnull() obtenemos un DataFrame con las mismas dimensiones que el original, pero con valores booleanos indicando si los registros son o no nulos (Figura 10).

FACULTAD	CAR_CODIGO	CARRERA	MAT_CODIGO	...	NIVEL	CREDITOS	NUMERO_VECES	CEDULA	NOMBRE
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False
False	False	False	False	...	False	False	False	False	False

Figura 10, Gráfico que muestra la cantidad de valores nulos.

La presencia de valores faltantes en cualquier conjunto de datos plantea un desafío al llevar a cabo cualquier tipo de análisis. Es necesario eliminar estos valores de cara a automatizar los procesos de análisis. En el *dataframe* se va proceder a eliminar las filas y columnas que contenga valores nulos. Con el comando: *df.dropna()* eliminamos las filas de nuestro conjunto de datos (Figura 11).

FACULTAD	CAR_CODIGO	CARRERA	MAT_CODIGO	...	NIVEL	CREDITOS	NUMERO_VECES
FACULTAD DE SALUD Y CULTURA FISICA	2	ENFERMERIA	1636	...	6	80	0
FACULTAD DE SALUD Y CULTURA FISICA	2	ENFERMERIA	1640	...	7	80	0
FACULTAD DE SALUD Y CULTURA FISICA	2	ENFERMERIA	1638	...	7	160	0
FACULTAD DE SALUD Y CULTURA	2	ENFERMERIA	1642	...	7	120	0

Figura 11, Gráfico sin valores nulos

4.2.6 Preparación y muestreo de los datos

A partir de los datos originalmente capturados, se van generar atributos derivados, o valores transformados de atributos existentes, en función de los requerimientos para preparar la entrada al análisis planteado.

Para la proyección y oferta académica procedemos a transformar las columnas de tipo *string* a numéricas de tal forma que nos permita realizar los cálculos respectivos. En este primer data set nuestras columnas objetivo son:

- TIPO_ALUMNO, se identifica si es un alumno para primer nivel (Nuevo), o es un alumno que ya tiene una matrícula anterior (Continuar).

Utilizamos el comando:

```
ta=pd.get_dummies(df['TIPO_ALUMNO'])
df_m=pd.concat([df,ta],axis=1)
```

- SI SE ABRIO, es la columna que identifica que paso con las materias en los periodos anteriores, si la materia se abrió coloca ‘S’, caso contrario ‘N’.

Utilizamos el comando:

```
df['ABIERTO'] = df['SI SE ABRIO'].apply(lambda x: 1 if x == 'S' else 0).
```

Para generar el horario se procede a definir las columnas objetivas con las siguientes sentencias de Python.

Suma de horas presenciales y sincrónicas.

```
df['HORAS']=df['PRESENCIALES_SEMANALES']+df['SINCRONICAS_SEMANALES']
```

Se define una sola Ciudad de estudio, con materias que no tenga valores nulos.

```
df=df[(df['SED_CODIGO']==1)&(df['HORAS']!=0)                    &
(df['MAT_CODIGO'].duplicated(keep='first'))                     &
(df['MAT_CODIGO'].isin(df['MAT_CODIGO'].to_list()))].
```

En la Figura 12, se muestra los datos procesados.

SED_CODIGO	MAT_CODIGO	MAA_NIVEL	PRESENCIALES_SEMANALES	SINCRONICAS_SEMANALES	CUPO_GRUPO	COMPONENTES	HORAS
1	1018	1	0	2	3	S	2
1	1026	1	0	3	10	P	3
1	1603	1	0	3	8	P	3
1	1288	1	0	2	12	S	2
1	1013	1	3	0	25	P	3
...	...	...	...	...	...	...	...
1	1644	8	1	0	102	P	1
1	1643	8	0	1	60	S	1
1	1647	9	1	1	47	P	2
1	1645	9	1	0	60	S	1
1	1645	9	1	0	60	P	1

Figura 12. Gráfico con columnas objetivo preparadas para construcción de horarios.

Por otro lado, se procede a preparar los datos para el distributivo, eliminando las filas *null*, y utilizando el método **isin** para verificar si el *data frame* contiene los valores especificados (Figura 13).

```
dp = pd.read_excel('materia_horarios.xlsx',sheet_name='docentes').dropna()
dp=dp[dp['MAT_CODIGO'].isin(df['MAT_CODIGO'].to_list())]
```

	SED_CODIGO	CAM_CODIGO	FAC_CODIGO	CAR_CODIGO	MAT_CODIGO	CODIGO_DOCENTE	NIVEL	CEDULA
0	1	1	17	2	1288	1347.0	1	9.612303e+08
1	2	10	17	2	1288	1347.0	1	9.612303e+08
2	2	10	17	2	1013	1274.0	1	1.756834e+09
3	1	1	17	2	1013	1274.0	1	1.756834e+09
4	2	10	17	2	1018	1936.0	1	9.594948e+08
...	...	...	...	...	...	...	...	...
395	2	10	17	2	1645	1720.0	9	9.630134e+08
396	2	10	17	2	1645	1352.0	9	9.620043e+08
397	2	10	17	2	1645	1352.0	9	9.620043e+08
398	1	1	17	2	1645	1795.0	9	1.725275e+09
399	1	1	17	2	1645	1795.0	9	1.725275e+09

Figura 13. Gráfico con columnas objetivo preparadas para distributivo académico.

Para la generación de horarios es importante definir las *constraints* o restricciones necesarias que permita definir un horario que se ajuste a los requerimientos de la IES, para este proyecto se ha tomado las que se ha considerado en la etapa de investigación.

- No más de 3 materias por hora y no menos de una (C1).

- Número de horas de la asignatura igual al número de horas esperadas (C2).
- Total, horas de una asignatura menor al máximo de horas de una materia en un día (C3).
- Identificar que días una materia se imparte (C4).
- Verificar que no se dé menos o más días de clases para una materia (C5).
- Dar las asignaturas en el menor número de dias (C6).
- Tener el menor número de materias en una hora (C7).

4.2.7 Realización del modelo

En el proyecto el problema es identificar la proyección, oferta y horario académico a partir de un conjunto de datos de entrenamiento que contiene observaciones cuya categoría de pertenecía es conocida, se utilizara el conjunto de datos de una IES anónima para crear un modelo que prediga la proyección de alumnos con la oferta académica en un periodo, se construyó 3 modelos diferentes utilizando diferentes algoritmos Decisión Tree, Support Vector Machines, Naive Bayes.

Después de construir cada modelo, los evaluaremos y compararemos qué modelo es el mejor para nuestro caso. Luego intentaremos optimizar nuestro modelo ajustando los hiperparámetros. Por último, guardaremos el resultado de la predicción de nuestro conjunto de datos y luego almacenaremos nuestro modelo para su reutilización.

Para comenzar, cargaremos algunas bibliotecas básicas como Pandas y NumPy y luego realizaremos algunas configuraciones.

Para la evaluación de los modelos se trabajará con validación cruzada, separando los datos de entrenamiento y datos de prueba, de tal manera que se pueda probar su eficiencia con un conjunto de datos no conocidos.

Estos datos se utilizarán en todos los modelos seleccionados de tal manera que los resultados obtenidos sean compatibles.

El proceso que se utiliza para la separación de la data se visualiza en la siguiente Figura 14.

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, classification_report

# Normalizacion de los datos
scaler = MinMaxScaler()
X = grouped_data[cols_group]
y = grouped_data['ABIERTO']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=False)

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Figura 14. Gráfico de datos de entrenamiento y prueba

4.2.7.1 Modelado con Arboles de decisión.

Se empieza el análisis para este algoritmo, utilizando la instrucción train_test_split procedemos a separar los datos de entrenamiento y prueba.

Posteriormente se define el algoritmo, entonces desde *skelarn.tree* se importa *DecisionTreeClasiffier*. Lo cual indica que desde el módulo de árboles de *skelearn*, se importa el algoritmo de árboles de decisión de clasificación.

Ahora se va definir el algoritmo y lo configuraremos para crear el modelo y entrenarlo.

Para entrenar el modelo con árboles de decisión no se ajustará el hiperparámetro de la profundidad del árbol, es decir el modelo avanzará hasta que no haya más bifurcaciones.

El objetivo es crear un modelo que prediga que variables influyen más en la determinación de una oferta y proyección académica (Figura 15).

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, classification_report

# Normalizacion de los datos
scaler = MinMaxScaler()
X = grouped_data[cols_group]
y = grouped_data['ABIERTO']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=False)

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
classifier = DecisionTreeClassifier()

classifier.fit(X_train_scaled, y_train)
y_pred = classifier.predict(X_test_scaled)
custom_confusion_matrix(y_test, y_pred)
```

Figura 15. Implementación de modelado con árboles de decisión.

4.2.7.2 Modelado con Support Vector Machine (SVC)

Las SVM (Máquinas de Vector Soporte) representan un algoritmo de clasificación y son consideradas entre los clasificadores más efectivos en diversas circunstancias. Fundamentado en el concepto de hiperplano, el SVM obtiene resultados destacados cuando la división entre clases es cercana a lineal. Sin embargo, su eficacia disminuye significativamente cuando dicha división no se asemeja a una línea.

La finalidad es entrenar un modelo SVM capaz de clasificar las observaciones, vamos a separar los datos de "X" y "y", también se va proceder a separar los datos de entrenamiento y prueba, para ello utilizaremos la instrucción de train_test_split, la cual nos facilita bastante este procedimiento.

A continuación, se establecen el algoritmo, en este contexto, instruiremos al programa, de *sklearn*.svm vamos a importar SVC, con esto ya podemos implementar este algoritmo dentro de nuestro programa.

Ahora vamos a entrenar el modelo utilizando los datos de entrenamiento separados anteriormente, con el modelo entrenado se realiza una predicción con los datos de prueba (Figura 16).

```
from sklearn.svm import SVC


# Assuming you have already split your data into X_train_scaled, X_test_scaled, y_train, y_test

# Create an SVM classifier
classifier = SVC()
classifier.fit(X_train_scaled, y_train)
y_pred = classifier.predict(X_test_scaled)
custom_confusion_matrix(y_test, y_pred)
```

Figura 16. Implementación de modelado con Support Vector Machine

4.2.7.3 Modelado con Naive Bayes

Naive Bayes es un modelo bayesiano, es decir, probabilístico, donde todas las predicciones se basan en calcular probabilidades, Naive *Bayes Gausiano* asume que los datos siguen una distribución *gausiana*.

Para la implementación de este algoritmo, primero se debe definir el módulo, *sklearn.naive_bayes*, e importar la clase que será *GaussianNB*. Este módulo cuenta con distintas clases, pero la que se utilizara en este proyecto es la más utilizada de todas. Una vez que se hace la definición, se procede al entrenamiento para esto se utiliza *fit*() de la mano con el algoritmo, para la realización de una predicción se usa *predict*(). Antes de llevar a cabo ambas directrices, se establecerán con anticipación las variables independientes y dependientes para que puedan ser empleadas.

Luego se realiza una predicción de conjunto a los datos de pruebas que separamos anteriormente (Figura 17).

```
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

clf = GaussianNB()

# Train the classifier on the training data
clf.fit(X_train_scaled, y_train)

# Make predictions on the test data
y_pred = clf.predict(X_test_scaled)

custom_confusion_matrix(y_test, y_pred)
```

Figura 17. Implementación de modelado con Naybe Bayes

4.2.7.4 Modelo de optimización

Por otro lado, para resolver el problema de los horarios se utilizará un modelo de optimización, a través de una colección de paquetes que en este caso es Pyomo. Para aplicar los resultados matemáticos y técnicas numéricas de la teoría de optimización, se hizo un delineamiento claro para los límites del sistema a optimizar, también se definió los parámetros cuantitativos que se utilizarán como criterio en base al cual serán clasificadas las alternativas para determinar la mejor opción.

Tabla 14. Variables del problema inicial.

Variables	Valores	Descripción
días	['l','m','x','j','v','s']	Días de la semana, se utiliza ‘x’ para miércoles, y domingo no se planifica clases.

horas	[f"h_{i}" for i in np.arange(7,22,1)]	Se define las horas de clase de 07:00 AM, 22:00 PM, cada registro será de 1 hora.
asignaturas	('SB_'+df['COMPONENTES']+df['MAT_CODIGO'].astype(str)).to_list()	Para las asignaturas se define los componentes Presencial y Sincrónico, junto con el número de horas de cada uno de ellos, mismos que vienen del plan de estudios.
horas_por_asignatura	dict(zip('SB_'+df['COMPONENTES']+df['MAT_CODIGO'].astype(str),df['HORAS']))	Horas por asignatura.
max_horas_por_dia	3	Máximo de horas a planificar por materia y por día.
model	pyo.ConcreteModel()	Inicializar la instancia del modelo matemático ConcreteModel()

Dentro de los parámetros se define:

- Horas por materia.
- Mínimo de días por materia.
- Máximo de días posible por asignatura.
- Cuantas asignaturas se da por día.
- Cuantas materias se da por hora.

Establecer el calendario, incorporando al modelo diversos parámetros y variables adicionales que nos permitan formular restricciones de manera efectiva.

Se define en adelante lo siguiente:

1. Parámetros, cualquier componente que tome un valor definido y fijo.
2. Variables, aquellos elementos cuyo valor están sujeto a las decisiones que adoptadas

A continuación, se detalla la decisión a lo que se va a restringir en el modelo matemático.

C1-No más de 3 materias por hora y no menos de una

Esta limitación establece que, en cada hora de cada día, se deben ofrecer exactamente tres asignaturas y no se permite menos de una.

C2-Número de horas de la asignatura igual al número de horas esperadas.

En este escenario, será suficiente asegurarnos de que el número total de asignaciones para cada materia sea precisamente igual a la cantidad de horas de clase por semana.

C3-Total horas de una asignatura < al máximo de horas de una materia en un día.

En esta situación, se espera que la suma total de horas de cada asignatura por día sea igual o inferior al límite máximo diario establecido.

C4-Identificar qué días una materia se imparte.

Esta limitación tiene la finalidad de activar la variable indicadora si la materia se enseña en ese día específico.

C5-Verificar que no se dé menos o más días de clases para una materia.

Se establecerá como requisito que cada asignatura sea impartida en un rango de días.

C6-Bloques consecutivos

Es crucial que las materias se enseñen en secuencia continua (horas consecutivas) debido a que no sería coherente ofrecer 1 hora de la materia A, luego 1 hora de la materia B y después otra hora de la materia A. Para lograr esto, se emplearán diversos métodos para modelar la restricción de manera precisa.

A menos que una materia pueda ocupar todo un día, nunca deberá asignarse la misma materia al final y al comienzo del día.

C7- Reducir la cantidad de días en los que se imparten las asignaturas.

Por lo general, será de interés impartir las materias en el menor número de días factible, aunque sin hacerlo una limitación rígida, para posibilitar la búsqueda de una solución del modelo incluso en situaciones desafiantes de ajuste. Para fomentar esta tendencia, se puede aplicar una penalización al total de días "extra" en los que se asignan materias, en comparación con la situación ideal donde cada materia se enseña durante el menor número posible de días (cifra que resulta del redondeo del número de horas de clase entre el máximo permitido por día).

El modelo de optimización, tiene como objetivo de cumplir con todas las limitaciones, también debe tener la capacidad de comparar soluciones ante distintas decisiones hasta identificar la óptima. Para lograr esto, se evaluará la programación semanal a partir de la suma de las preferencias asignadas (con un valor predeterminado de 1 para todas las asignaciones) y se impondrán penalizaciones por cada día adicional utilizado en comparación con el mínimo requerido para impartir cada materia.

Después de haber formulado el problema utilizando Pyomo, llega el momento de transmitirlo al solucionador y verificar si puede encontrar una solución viable. Además, es necesario asegurarse de que se respeten todas las asignaciones de horarios y que no existan casos donde las materias estén programadas en un mismo día en bloques horarios no contiguos.

4.2.8 Evaluación de los resultados

En este punto vamos a verificar que tan preciso fueron los modelos desarrollados, utilizó el módulo de métricas que proporciona la librería *scikit learn* para el efecto.

A continuación, se revisará la matriz de confusión de los diferentes modelos desarrollados.

- **Árboles de Decisión**

Dentro de la construcción del modelo habíamos implementado la matriz de confusión para esto se importó de *metrics* de *sklean* el método *confusión_matrix* y se implementa con los datos predichos y datos reales.

```
Specificity: 0.6052631578947368
Accuracy: 0.6106194690265486
Confusion Matrix:
 [[69 45]
 [43 69]]
```

Figura 18. Matriz de confusión – Arboles de Decisión.

Como es evidente, los datos en la diagonal principal corresponden a las predicciones acertadas, mientras que en la diagonal secundaria se encuentran los errores. Por consiguiente, al sumar estos elementos, obtenemos un total de 138 predicciones correctas y 88 predicciones incorrectas.

Para ver la precisión del modelo, se importó "*precision_score*" de *metrics* y se implementó junto con los datos predichos y los reales.

```
Precisión del modelo:
0.603448275862069
```

El resultado de este cálculo es de 0.60, este es un valor no aceptable.

- **Support Vector Machine**

Se procede a comparar los valores obtenidos luego de realizar la predicción con los datos reales para ver que tal ha sido el modelo que hemos construido, pero esta vez vamos a aplicar las métricas con las que cuenta los algoritmos de *Machine Learning*, se procede a verificar la matriz de confusión. El resultado obtenido el que se observa.

```
Specificity: 0.7982456140350878
Accuracy: 0.6238938053097345
Confusion Matrix:
 [[91 23]
 [62 50]]
```

Viendo este resultado ya se puede intuir que el modelo ha sido correcto, pero es necesario analizar la precisión del modelo.

```
Precisión del modelo:
0.684931506849315
```

Para el cálculo se importó la función precisión y se aplicó junto con los datos de "y" de prueba y los obtenidos en la predicción. El resultado obtenido es 0.69. Por lo que consideramos que el modelo cumple con su función.

- **Support Naive Bayes**

Vamos a verificar como es el modelo utilizando las métricas de los problemas de clasificación, para ello vamos a comenzar obteniendo la matriz de confusión. Para esto importamos del módulo *scikitlearn.metrics, confusion_matrix,* y aplicamos esta instrucción junto a los datos de prueba y los obtenidos en la predicción realizada previamente.

```
Specificity: 0.2631578947368421
Accuracy: 0.5309734513274337
Confusion Matrix:
 [[30 84]
 [22 90]]
```

El resultado es que tenemos 120 datos predichos correctamente y 106 datos erróneos obtenido luego de realizar la predicción.

Viendo este resultado podemos concluir que el modelo no predijo la gran mayoría de los datos por lo que no es un buen modelo que podamos utilizar.

Ahora confirmemos la precisión del mismo, para esto importamos *precision_score* del módulo *sklearn.metrics* y lo implementamos de igual forma junto a los datos de entrenamiento y los predichos.

```
Precisión del modelo:
0.5172413793103449
```

Ahora veamos la precisión del mismo, para esto importamos *precision_score* del módulo *sklearn.metrics* y lo implementamos de igual forma junto a los datos de entrenamiento y los predichos, el valor es muy bajo, lo que indica que no se puede aplicar el modelo.

- **Modelo de optimización**

Se utilizó Pyomo para resolver el problema de modelado y optimización de horarios, el modelo de optimización utilizado consta de variables de decisión, restricciones y un objetivo de optimización, también podemos observar que en ningún caso existen bloques no consecutivos asignados para la misma asignatura en el mismo día.

DIA HORA	Lunes	Martes	Miercoles	Jueves	Viernes	Sabado
07:00	Materia 1623 Presencial Docente 1934	Materia 1623 Presencial Docente 1934	Materia 1618 Presencial Docente 1719			
08:00	Materia 1623 Presencial Docente 1934		Materia 1618 Presencial Docente 1719	Materia 1606 Presencial Docente 1719		
09:00				Materia 1606 Presencial Docente 1719		
10:00				Materia 1606/1638 Presencial/Sincronico Docente 1719/9835		
11:00				Materia 1638 Sincronico Docente 1719		
12:00				Materia 1638 Sincronico Docente 1719		
13:00						
14:00						
15:00						
16:00						Materia 1606 Presencial Docente 1719
17:00						Materia 1606 Presencial Docente 1719

Figura 19. Gráfico de horario académico generado.

Como se puede apreciar en el grafico generado, la restricción definida está funcionando, se puede tomar la materia Materia-Presencial 1618 como ejemplo, se puede apreciar que los

bloques de clases no son consecutivos, se busca dar la asignatura en el menor número de días posibles, y bien definidas la cantidad de horas y días en el horario.

5. Conclusiones y Recomendaciones

5.1. Conclusiones

- Se utilizó la metodología CRISP-DM para entender las tareas necesarias de este proyecto de ciencia de datos, logrando así deducir los criterios para medir el éxito en el proyecto, también permitió establecer un primer contacto con el problema, y en la fase de preparación de los datos se encontró una relación directa con la técnica de modelado.
- Teniendo en cuenta los objetivos planteados para la realización de este trabajo, el principal aporte logrado es el entendimiento de la oferta, distributivo, y horario académico que una institución de educación superior debe realizar, que puede servir como referente para trabajos posteriores y para la consulta de investigaciones que se vayan a dar con respecto a este tema.
- Las tecnologías asociadas al enfoque de Ciencia de Datos ya han comenzado a tomar madurez y se vislumbran grandes oportunidades y retos en su utilización, optimización y adaptación a diferentes dominios de datos. Sin embargo, ya se encuentran resultados que muestran sus beneficios en aspectos como la reducción de tiempos, optimización de recursos y mayor flexibilidad.
- Intentar llegar a una predicción de datos en cuanto a que materias se puede abrir, con que docentes y en que horario, es factible, pero siempre dependerá de factores externos y políticas dadas por cada IES, como es, carga horaria de docentes, espacios físicos, y uno principal el factor económico.

5.2. Recomendaciones

- Se entrega en este trabajo los tres modelos de clasificación ajustados, adicional el algoritmo de optimización de horarios; sin embargo, se recomienda trabajar más en el perfeccionamiento del algoritmo de optimización, de acuerdo al modelo de negocio de cada institución, por las restricciones definidas en nuestro caso.
- Los modelos de datos son de gran ayuda al momento de analizar este tipo de proyectos, no obstante, se debe tener cuidado al momento de realizar la limpieza de datos, dado que se debe considerar las variables que estamos exportando para definir nuestras variables objetivas, en este caso se recomienda crear una propia

estructura de base de datos, que permita mediante consultas personalizadas, hacer limpieza de datos y revisar la calidad de la información.

6. Bibliografía

CES. (23 de Abril de 2019). https://www.ces.gob.ec/lotaip/Anexos%20Generales/a3_Reformas/r.r.academico.pdf

coin-or.org. (2005). *CBC*. CBC: https://www.coin-or.org/Cbc/cbcuserguide.html#id3342315

Fallarino, N., Leite, A., & Cremades, R. (2020). Estudio de caso sobre el desarrollo de la competencia oral en Educación Infantil y Primaria en un centro público. *Red de Información Educativa, 31*(03), 10. https://redined.educacion.gob.es/xmlui/bitstream/handle/11162/200417/Fallarino.pdf?sequence=1&isAllowed=y

Guerrero, A. (2018). Reglamento de Régimen Académico de la Universidad Nacional de Chimborazo, reformado. *UNACH*, 46. https://www.unach.edu.ec/images/reglamentos/regl-reg-acad.pdf

Hernández, J., & Martínes, F. (27 de Enero de 2019). Topografía de las características de la IA relevantes para la seguridad. *Repositorio de la Universidad Politécnica de Valéncia*, 9. https://doi.org/http://hdl.handle.net/10251/14656

Huancas, J., & Vargas, J. (2021). Desarrollo de un modelo de procesos para la construcción de software en microempresas peruanas desarrolladoras de software. Caso de estudio: Sistema Inteligente ERP SAC. *Repositorio de la Universidad Señor de Sipán*, 208. https://repositorio.uss.edu.pe/bitstream/handle/20.500.12802/9129/Huancas%20Montenegro%20Jeiner%20%26%20Vargas%20Moreno%20Jorge.pdf?sequence=1&isAllowed=y

Oviedo, A., & Jiménez, J. (Julio de 2019). MINERÍA DE DATOS EDUCATIVOS: ANÁLISIS DEL DESEMPEÑO DE ESTUDIANTES DE INGENIERÍA EN LAS PRUEBAS SABER-PRO. *Revista Politécnica, 15*(29), 14. https://doi.org/https://doi.org/10.33571/rpolitec.v15n29a10

Pando, A., & Zarate, W. (2021). Aplicación de un modelo de minería de datos para identificación de patrones que influyen en la deserción académica en el instituto superior Leonardo Davinci usando IBM SPSS modeler y la metodología CRISP-DM. *Repositorio Digital de la Universidad Privada Antenor Orrego*, 13. https://doi.org/https://hdl.handle.net/20.500.12759/7033

Paredes , J., Bonilla , F., & López , E. (2020, Julio 15). Análisis de la pertinencia de la oferta académica vigente y vigente planificada del Instituto Superior Tecnológico La Maná. *Revista Académica y Científica VICTEC, 1*(1). https://server.istvicenteleon.edu.ec/victec/index.php/revista/article/download/7/5/13

Pyomo. (2021). *pyomo.org*. http://www.pyomo.org/about

Ramos, J. (2019). Aprendizaje automático para flujos de datos. 156. https://oa.upm.es/56025/1/TFM_JAVIER_RAMOS_FERNANDEZ.pdf

Salazar, A., Villarreal, H., & Mosquera, J. (2022). Herramientas y prácticas para el apoyo en la toma de decisiones en proyectos de migración de datos. *Repositorio UNIAJC*, 70. https://repositorio.uniajc.edu.co/handle/uniajc/1311

Schab, E. A., De Battista, A. C., Cagnina, L. C., & Herrera, N. E. (2019). Descubrimiento de conocimiento en bases de datos. *Repopsitorio Institucional CONICET digital*. https://ri.conicet.gov.ar/handle/11336/161266

Tingo, F., Chavez, S., Cevallos, M., Soria, R., & Yépez, J. (2018, Agosto). Prácticas sociales que inciden en el rendimiento académico y obtención de diplomas de Bachillerato Internacional (BI) en estudiantes de instituciones educativas fiscales: Estudio de casos. https://educacion.gob.ec/wp-content/uploads/downloads/2019/02/practicas-sociales-diploma-internacional.pdf