VOLUMEN 6, N°2 / JULIO - DICIEMBRE 2024 / e - ISSN: 2661

ANÁLISIS DE DATOS PARA LA OPTIMIZACIÓN EFICIENTE DE HORARIOS Y APRENDIZAJE AUTOMÁTICO

Data analysis for efficient schedule optimization and Machine Learning

Rogelio Escobedo Mitre [1]	isc.rogelio.em@tectijuana.edu.mx
Ángeles Quezada Cisnero [2]	angeles.quezada@tectijuana.edu.mx
Bogart Yair Marquez Lobato [3]	bogart@tectijuana.edu.mx
Arnulfo Alanis Garza [4]	alanis@tectijuana.edu.mx

[1, 2, 3, 4] Tecnológico Nacional de México (Instituto Tecnológico de Tijuana)

RESUMEN

En los últimos años, la integración de técnicas de aprendizaje automático (AP), en los sistemas de gestión escolar ofrece varias oportunidades para mejorar la eficiencia y la toma de decisiones en el ámbito educativo. Al aplicar el AP en la educación, se pueden obtener beneficios significativos. Sin embargo, es importante tener en cuenta que la integración exitosa de técnicas de AP en los sistemas de gestión escolar requiere una infraestructura de datos sólida, la recopilación adecuada de datos y la consideración de cuestiones éticas y de privacidad. Además, el AP no debe reemplazar la interacción humana en la educación, sino complementarla y mejorarla, brindando a los educadores y estudiantes herramientas adicionales para el éxito educativo. Debido al incremento en la retícula de la carrera de Ingeniería en Sistemas Computacionales es necesario llevar a cabo una predicción más confiable y de manera automática de los horarios por semestre. Para abordar el problema de generación de horarios de manera manual, se llevó a cabo un análisis exhaustivo del proceso existente. Esto implica recopilar información relevante sobre cómo se realiza actualmente la generación de horarios por semestre en la institución educativa. Se estudio el enfoque actual utilizado para generar los horarios, además analizar los problemas y las limitaciones asociadas al proceso manual. Se investigarán diferentes técnicas de Machine Learning que podrían aplicarse al problema de generación de horarios. Esto podría incluir algoritmos de optimización, algoritmos de agrupamiento o clasificación, algoritmos genéticos u otros enfoques de aprendizaje automático que puedan adaptarse al problema específico.

Palabras Clave: aprendizaje automático, automatización, horarios.

ABSTRACT

In recent years, the integration of machine learning (ML) techniques into school management systems offers several opportunities to enhance efficiency and decision-making in the educational field. Applying ML in education can yield significant benefits. However, it is important to note that the successful integration of ML techniques into school management systems requires a robust data infrastructure, proper data collection, and consideration of ethical and privacy issues. Furthermore, ML should not replace human interaction in education but rather complement and improve it by providing educators and students with additional tools for educational success. Due to the increased complexity in the curriculum of the Computer Systems Engineering program, it is necessary to carry out a more reliable and automated prediction of semester schedules. To address the manual scheduling generation problem, a comprehensive analysis of the existing process was conducted. This

Fecha de Recepción: 23/nov/2023. **Fecha de Aceptación:** 01/mar/2024 **DOI:** https://doi.org/10.47187/perspectivas.6.2.223

involved gathering relevant information on how semester schedules are currently generated in the educational institution. The current approach used for scheduling was studied, along with an analysis of the problems and limitations associated with the manual process. Various ML techniques that could be applied to the scheduling generation problem were investigated. This could include optimization algorithms, clustering or classification algorithms, genetic algorithms, or other machine learning approaches that can be adapted to the specific problem.

Keywords: Machine learning, automation, schedules.

I. Introducción

En los últimos años, el avance de la tecnología ha transformado muchos aspectos de nuestra sociedad, y el ámbito educativo no ha sido una excepción. La integración de técnicas de Aprendizaje Automático (Machine Learning) en los sistemas de gestión escolar ha surgido como una solución prometedora para mejorar la eficiencia y la toma de decisiones en las instituciones educativas [1] [2].

En la actualidad, las instituciones educativas se enfrentan a desafíos constantes en la gestión y organización de sus actividades académicas. El seguimiento y análisis manual de datos estudiantiles, la planificación de horarios y la adaptación de los programas educativos a las necesidades individuales de los estudiantes son tareas que demandan un tiempo considerable y pueden resultar propensas a errores.

A su vez la falta de atención en la programación de horarios en instituciones de educación básica y media, resalta la importancia de considerar cuándo es más propicio el aprendizaje, a diferencia de la programación de horarios universitarios, la programación de horarios escolares se centra en las clases y no en los estudiantes [17].

En la literatura científica y en la industria, se han realizado avances significativos en el desarrollo de sistemas de gestión escolar basados en Aprendizaje Automático. Algunos estudios se han centrado en la predicción del rendimiento académico de los estudiantes, utilizando algoritmos de regresión y clasificación para identificar factores que influyen en el éxito o fracaso escolar. Otros trabajos han abordado la optimización de la planificación de horarios, aplicando algoritmos de agrupamiento y programación lineal para asignar eficientemente recursos y minimizar conflictos [3] [4].

Además, se han desarrollado sistemas de recomendación de asignaturas y cursos, utilizando técnicas de filtrado colaborativo y análisis de contenido para sugerir opciones académicas adaptadas a las preferencias y habilidades de cada estudiante [5] [6].

En universidades como Waterloo han desarrollado un sistema que se basa en una filosofía "dirigida por la demanda", donde los estudiantes eligen primero sus cursos y el sistema intenta encontrar el mejor horario para maximizar el número de solicitudes satisfechas donde el problema se descompone en subproblemas manejables que se resuelven secuencialmente utilizando una heurística ávida para asignar horarios a secciones y un algoritmo de relajación lagrangiana para asignar aulas, sin embargo la asignación de profesores no se hace automática a las secciones de los cursos; ese proceso generalmente se realiza manualmente con antelación [18].

Se sabe, con antelación, que la situación de horarios automáticos no es una situación nueva, y tampoco algo esporádico, ya que universidades como Don Bosco ya que es sabido que el proceso de elaboración de horarios académicos sigue siendo realizado de manera manual, ya que personal docente y administrativo se reúne previo al inicio del período a iniciar para analizar en detalle las estadísticas relacionadas con la población estudiantil, el cuerpo docente y las instalaciones disponibles. Este análisis tiene como objetivo principal la creación de horarios que permitan una gestión eficiente de los recursos institucionales [19].

La programación de horarios académicos plantea un desafío particular que reside en la problemática más amplia de la asignación de recursos. En la comunidad científica, se conoce este desafío

como el "Problema de Programación de Horarios Universitarios". Este tipo de problemas implica la generación de horarios para tareas específicas, con el objetivo de cumplir de manera óptima con condiciones y requisitos particulares. A lo largo del tiempo, se han abordado estos problemas de programación de horarios utilizando diversos enfoques, tales como algoritmos basados en la Colonia de Hormigas, Búsqueda Tabú, Coloreo de Grafos y Algoritmos Genéticos, entre otros [20].

En este contexto, el presente proyecto de análisis de datos para la optimización eficiente de horarios y aprendizaje automático tiene como objetivo abordar estas problemáticas para ofrecer una solución integral y adaptada a la carrera de Ingeniería en Sistemas Computacionales, que permita optimizar la gestión académica con respecto a la cantidad de materias abiertas por semestre para mejorar la oferta académica para estudiantes de tal forma que se brinde un mejor rendimiento en el avance reticular. A través de un enfoque innovador y personalizado, se busca establecer una base sólida para la transformación de las prácticas educativas y sentar las bases para futuros avances en la gestión escolar impulsada por la inteligencia artificial

II. Conceptos relacionados

II. I. Aprendizaje Automático

El aprendizaje automático, conocido como Machine Learning en inglés (ML), es una disciplina dentro del campo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos capaces de aprender y mejorar automáticamente a través del análisis de datos. Estos algoritmos tienen la capacidad de identificar patrones complejos y realizar predicciones o tomar decisiones basadas en ellos, sin requerir una programación específica para cada tarea [7]. El aprendizaje automático se basa en principios estadísticos y computacionales, utilizando técnicas que permiten a las máquinas procesar y analizar grandes volúmenes de datos con el fin de extraer información relevante. Al utilizar estos datos, los algoritmos pueden descubrir relaciones y regularidades ocultas, lo que les permite aprender y adaptarse a medida que se les suministra más información.

Ciclo de vida del ML. Existe un ciclo básico para poder realizar proyectos de machine learning, como se muestra en la Figura 1.

Figura 1
Componentes de ML

Fuente: *Elaboración propia (imagen de referencia tomada de https://keepcoding.io/blog/ciclo-de-vida-de-un-proyecto-en-machine-learning/)*

II. II. Principales algoritmos de ML

Algoritmo de aprendizaje supervisado

El algoritmo de aprendizaje supervisado es una técnica de ML en la que se utilizan conjuntos de datos etiquetados para entrenar un modelo y realizar predicciones sobre nuevos datos no etiquetados. Este tipo de algoritmo se basa en la idea de que existe una relación entre las características o variables de entrada y una variable de salida deseada, y el objetivo es aprender esta relación a partir de los datos de entrenamiento [9].

Los algoritmos más utilizados en el aprendizaje supervisado incluyen árboles de decisión, clasificación de Naive Bayes Ingenuo, regresión por mínimos cuadrados, regresión logística, métodos de ensamble y Máquinas de Soporte Vectorial por sus siglas en inglés SVM (support vector machines). Cada algoritmo tiene sus propias ventajas y desventajas, y su elección dependerá del problema específico y los datos disponibles.

Árboles de decisión:

Los árboles de decisión son ampliamente utilizados en el análisis de datos y la toma de decisiones automatizada, ya que son fáciles de interpretar y pueden manejar conjuntos de datos grandes y complejos. Además, son una base importante para algoritmos más avanzados, como Random Forests y Gradient Boosting, que combinan múltiples árboles de decisión para mejorar su rendimiento

predictivo. [11]

Clasificador Bayesiano Ingenuo (Naive Bayes):

Es un algoritmo de clasificación en el campo del aprendizaje automático y la minería de datos. Este clasificador se basa en el teorema de Bayes y asume que las características (o variables) que se utilizan para la clasificación son independientes entre sí, lo que se conoce como una suposición "ingenua (naive)" [12].

Regresión por mínimos cuadrados:

Es un método estadístico utilizado para modelar la relación entre una variable dependiente (o respuesta) y una o más variables independientes (o predictores) en un conjunto de datos. El objetivo principal de la regresión por mínimos cuadrados es encontrar una ecuación o modelo que minimice la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo. En otras palabras, se busca ajustar una línea o una superficie de manera que se minimice la suma de los errores cuadrados. [13]

Regresión logística:

Es un método estadístico y un modelo de regresión utilizado para analizar la relación entre una variable binaria dependiente (es decir, una variable que toma dos valores, típicamente 0 y 1) y una o más variables independientes. A diferencia de la regresión lineal, que se utiliza para predecir valores continuos, la regresión logística se emplea para problemas de clasificación, como la predicción de la probabilidad de pertenecer a una de las dos categorías [14].

III. Propuesta

Debido al avance de la tecnología el desarrollo de sistemas de gestión escolar ha sido una respuesta a la necesidad de optimizar la administración y el manejo de la información en las instituciones educativas. Tradicionalmente, estas tareas se realizaban de manera manual y consumían una gran cantidad de tiempo y recursos. Sin embargo, con el avance de la tecnología, se han desarrollado soluciones informáticas que permiten automatizar y simplificar estos procesos.

En cuanto al desarrollo del contexto de intervención, es importante mencionar que existen diferentes sistemas de gestión escolar disponibles en el mercado. Sin embargo, la implementación de técnicas de aprendizaje automático (machine learning) en este tipo de sistemas es una tendencia emergente que ofrece ventajas adicionales.

El Análisis de datos propuesto para la optimización eficiente de horarios utilizará técnicas de aprendizaje automático para analizar y procesar los datos relacionados con la gestión académica y administrativa de una institución educativa.

El proyecto de Análisis de datos para la optimización eficiente de horarios y Aprendizaje Automático busca abordar esta problemática, ofreciendo una solución innovadora y eficiente para mejorar la gestión y el rendimiento de las instituciones educativas. Al aprovechar las capacidades del machine learning, se espera que el sistema pueda proporcionar análisis predictivos y recomendaciones basadas en datos, permitiendo a las instituciones educativas tomar decisiones informadas y mejorar su eficacia.

En resumen, el contexto de intervención del proyecto se enmarca en la necesidad de modernizar y optimizar la gestión escolar a través del uso de técnicas de machine learning. Con el objetivo de superar los desafíos y limitaciones de los sistemas tradicionales, el proyecto busca ofrecer una solución innovadora que mejore la eficiencia, la toma de decisiones y el rendimiento general de las instituciones educativas.

Se llevarán a cabo diversas actividades durante el desarrollo del proyecto. Estas actividades incluirán la recopilación y análisis de datos relevantes, la identificación de los requerimientos específicos de nuestra institución educativa, específicamente en el departamento de Sistemas y Computación y de la carrera de Ingeniería en sistemas computacionales, la creación de algoritmos de machine learning adaptados a las necesidades del sistema de gestión escolar, y la implementación

de un entorno de prueba para validar y ajustar la solución propuesta.

IV. Metodología

En la Figura 2 se describe el proceso metodológico, llevado a cabo en esta investigación, el cual consiste en los siguientes pasos: definición de la muestra, obtención de datos, análisis de datos y presentación de los datos.

Figura 2
Proceso metodológico de la investigación

Fuente: *Elaboración propia.*

IV.I. Definición de la muestra

En este proyecto, se utilizó un muestreo estratificado para seleccionar los periodos que formaron parte del estudio diagnóstico. La muestra fue seleccionada considerando la

información con la que se cuenta con respecto a las bases de datos almacenadas, considerando también diferentes estratos relevantes para el estudio, como, por ejemplo, semestre cursado del alumnado, periodo de ingreso del mismo (enero-junio o agosto-diciembre), rendimiento académico, entre otros.

El proceso de selección de la muestra fue ejecutado de manera cuidadosa. En particular, se puso un énfasis especial en la información contenida en las bases de datos institucionales, contenidas en el sistema escolar desde 11 años atrás y las bases de datos almacenadas y procesadas en archivos de Excel, las cuales se convirtieron en una fuente invaluable de datos para llevar a cabo este estudio.

Uno de los principales criterios utilizados para la estratificación de la muestra fueron los semestres cursados por el alumnado, tomando como base las cantidades de materias ofertadas por semestre y la cantidad de alumnos que tomaron dichas materias. De esta manera, se aseguró que cada estrato de semestre estuviera representado en la muestra en proporción a su importancia relativa en la población estudiantil.

Además, se consideró el periodo de ingreso de los estudiantes, dividiéndolo en dos categorías: aquellos que ingresaron en el primer semestre del año (enero-junio) y aquellos que lo hicieron en el segundo semestre (agosto-diciembre). Esta división se basó en la suposición de que el contexto académico podría variar según el momento en que los estudiantes comenzaron su formación, lo que podría influir en su rendimiento y en sus experiencias educativas.

El rendimiento académico de los estudiantes también se convirtió en un criterio de estratificación crucial. Se categorizaron los estudiantes en función de su desempeño académico, considerando aspectos como el promedio de calificaciones, la tasa de aprobación y otras métricas relacionadas. Esta estratificación permitió analizar si existían diferencias significativas en el rendimiento de los estudiantes en función de sus logros previos.

IV.II. Obtención de datos

El análisis de datos es una parte crucial en la toma de decisiones informadas, y la calidad de la información utilizada es esencial para obtener resultados precisos. En este contexto, la información utilizada se obtuvo de múltiples fuentes, principalmente de fuentes primarias. La fuente principal de datos en este caso fue el Sistema de Información Escolar (SIE). Este sistema, que actúa como un repositorio central de información relacionada con cuestiones académicas y de planificación educativa, proporcionó una base sólida para el análisis.

Un aspecto importante a destacar es el acceso directo a los horarios almacenados en el sistema SIE. Esto permitió una recopilación de datos eficiente y precisa, ya que los horarios son componentes esenciales para el análisis, especialmente en un entorno educativo. El acceso a estos datos en tiempo real y de manera directa garantizó la integridad y actualidad de la información utilizada.

Además, para contar con un panorama más completo y para rastrear tendencias y patrones a lo largo del

tiempo, se aprovecharon los archivos de Excel que contenían el historial de horarios recopilados y actualizados semestre tras semestre. Estos archivos históricos proporcionaron una perspectiva a largo plazo y permitieron evaluar la evolución de los horarios y su impacto en la planificación educativa.

En resumen, el análisis de datos se basó en una sólida combinación de fuentes primarias, incluyendo el sistema SIE, la colaboración del departamento de servicios escolares, el acceso a los horarios en tiempo real y el uso de archivos históricos en formato Excel.

A continuación, se presenta una representación de los horarios correspondientes al periodo académico AGO-DIC 2023, junto con la estructura utilizada por el sistema para generar y formatear los informes que sirven como fundamento para su posterior archivo en formato Excel, como se muestra en la figura 2.

Figura 3
Concentrado de materias, indicando el catedrático asignado, así como grupo, horario y aula

REAL	GRUPO	MATERIA	CR	PLAN	PAQ.	CATEDRATICO	NECES	OCUPAC	REP	LUNES HORA/AULA
	AED-1285SC1A	FUND DE PROGRAMACION	5.00	161	01A			000/000	0	10001100/91L1
	ACF-0901	CALC. DIFERENCIAL	5.00	161	01A			000/000	0	09001000/91L1
	ACC-0906	FUNDAM D INVESTIG	4.00	161	01A			000/000	0	12001300/91L1
	ACA-0907	TALL DE ETICA	4.00	161	01A			000/000	0	08000900/91L1
	AEF-1041SC1A	MATEM DISCRETAS	5.00	161	01A			000/000	0	11001200/91L1

Fuente: *Elaboración propia.*

IV.III. Obtención de datos

La información se ha sometido a un proceso de organización, con el objetivo de crear una estructura coherente que facilite el análisis. Esta organización se ha llevado a cabo utilizando diversas herramientas y técnicas, como tablas, hojas de cálculo y otros formatos apropiados para asegurar que los datos sean de fácil acceso y comprensión.

Para profundizar en el análisis, se realizó un estudio de correlación entre las variables presentes en el conjunto de datos. Esta fase fue esencial para identificar las relaciones y dependencias entre las variables. Para lograrlo, se emplearon técnicas avanzadas, como la matriz de correlación y gráficos de dispersión, que proporcionaron una visión más clara de cómo las variables se relacionan entre sí.

Por último, se llevaron a cabo comparaciones minuciosas entre grupos o categorías de datos para identificar diferencias significativas. Estas comparaciones se llevaron a cabo siguiendo un enfoque estadístico sólido, lo que permitió obtener conclusiones fundamentadas y relevantes.

En la figura 2 se muestra la información de grupos por semestre y periodo, en la que se pueden observar las cantidades de grupos creadas a partir de la información recopilada:

Figura 4
Tabla comparativa del historial de materias abiertas por semestre a lo largo de un periodo 7 años.

semestre	2016	2017	2018	2019	2020	2021	2022
1	3+r	4+r	4+r	3+2r	3+2r	4+3r	5
2	4+r	4+r	4+r	3+2r	3+2r	4+2r	4+r
3	2+r	2+r	3+r	3	3	3+1r	4
4	4	2	3	3	3	3+2r	4
5	3	2	2	2	2	3	3
6	3	2	3	2	2	3	3
7	2	2	2	2	2	2	3
8	2	2	2	2	2	3	2+r
9	1	1	1	2	2	2	2

V. Presentación de los datos

Para poder visualizar mejor la información y poder hacer un análisis estadístico se generaron tablas en Excel, así como gráficas para poder visualizar de forma clara y precisa dicha información generada durante el análisis de datos.

Durante la fase de presentación de resultados, se ha puesto un fuerte énfasis en la creación de informes que destaquen de manera clara y concisa los hallazgos esenciales obtenidos a través del análisis. Estos informes se han estructurado de manera que sean fácilmente accesibles y comprensibles para los interesados y partes involucradas en el proyecto.

En particular, para mejorar la visualización de los datos y permitir un análisis estadístico más efectivo, se ha llevado a cabo la creación de tablas en formato Excel. Estas tablas han sido diseñadas con precisión y han sido organizadas de manera lógica, lo que facilita la revisión de los datos y la realización de cálculos estadísticos cuando es necesario.

Además, se han creado gráficos y visualizaciones adecuadas para representar visualmente la información generada durante el análisis de datos. Estas representaciones visuales permiten una comprensión más rápida e intuitiva de los patrones, tendencias y relaciones presentes en los datos. Se ha prestado especial atención a la claridad y la precisión en la presentación de estos gráficos, asegurando que sean herramientas efectivas para comunicar los resultados a las partes interesadas.

En resumen, se ha realizado un esfuerzo significativo en la generación de informes que simplifican la comunicación de los hallazgos clave, aprovechando gráficos, tablas y visualizaciones apropiadas. El objetivo final es proporcionar una plataforma sólida para la toma de decisiones basadas en datos y garantizar que los resultados del análisis sean fácilmente accesibles y comprensibles para todos los involucrados.

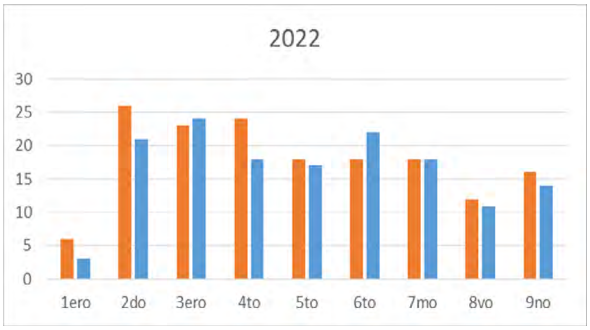
La información con la que se trabajó para poder realizar el proceso de análisis y evaluación de los periodos desde 2012 se obtuvo del programa SIE y se trabajó directamente en una hoja de cálculo donde se ordenó la información de tal forma que se pueda visualizar y trabajar para próximos semestres.

Primero, se debe observar cuántos grupos en total del primer semestre del periodo inmediato anterior. Enseguida se evalúa el segundo semestre del mismo periodo para analizar cuántos posibles repetidores serán candidatos a tomar las mismas materias del semestre en el que se requiere trabajar. Posteriormente se evalúa el segundo semestre del año inmediato anterior de tal forma que se pueda ver el avance y las diferencias/similitudes que se pueden presentar, tomando en cuenta el periodo de ingreso de los jóvenes (ingreso directo en agosto, o ingreso posterior al semestre propedéutico).

Después de analizar los periodos de cada semestre se muestra la cantidad de grupos, por semestre, de 3 periodos consecutivos, que son los que se requiere analizar para poder tomar las decisiones pertinentes para el siguiente semestre, como se muestra en la Figura 4.

La gráfica nos muestra la cantidad de grupos por semestre que se proyectaron en el periodo enero-junio y agosto diciembre 2022, con esta información se genera la proyección de grupos para el semestre enero-junio 2023, iniciando el primer semestre hasta el noveno. Con esta información se puede proyectar la cantidad de grupos que deberán ofertarse tomando en consideración los grupos de alumnos que estén en situación normal en cuanto a su carga académica, así como alumnos que tengan alguna materia reprobada.

Figura 5
Comparativo de cantidad de materias abiertas por semestre en el año 2022



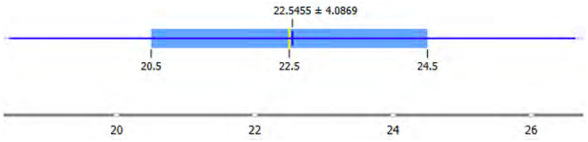
V. Resultados

Posterior al análisis y procesamiento de la información los datos obtenidos de la sección anterior se procedieron a utilizar herramientas como Box Plot para generar diagramas de caja y bigotes, ya que con esta herramienta gráfica nos permite visualizar la distribución y las estadísticas resumidas de un conjunto de datos. por lo tanto, es muy útil para identificar la simetría, la dispersión y la presencia de valores atípicos en un conjunto de datos de una manera visual y puede proporcionar una comprensión rápida de la distribución de los datos sin necesidad de una representación detallada de cada punto de datos.

Para lograr una proyección más precisa de la cantidad de materias a abrir por semestre, se utilizó la herramienta de Box Plot, donde se analizó la trayectoria desde agosto de 2012 hasta enero de 2023, abarcando 22 semestres consecutivos, es decir, un historial de 11 años. Esto nos permite obtener un promedio de la cantidad de materias a abrir por semestre.

Con los datos recopilados para el segundo semestre, se obtuvo que existe un rango mínimo de 14 grupos y un máximo de 32 grupos. La herramienta utilizada arroja un promedio de 22.5 grupos, con una variación de más o menos 4 grupos por período, como se muestra en la figura 4.

Figura 6
Diagrama caja y bigotes



Posteriormente se utilizó la herramienta Python para el desarrollo de un módulo que muestre en pantalla el resultado del análisis de la información proporcionada. El despliegue de la plataforma se inició con la incorporación de las librerías esenciales. La incorporación de Pandas, Random Forest y Numpy, conocidas por su profunda capacidad y funcionalidad, sentaron las bases sólidas sobre las cuales se construyó el marco de trabajo. Estas bibliotecas desempeñan un rol de importancia crítica en las fases de procesamiento, análisis y modelado de los datos, asegurando la robustez y efectividad de todo el sistema.

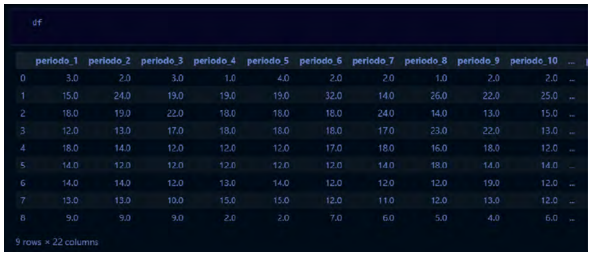
La subsiguiente etapa implicó la transferencia de datos al sistema. Se procedió a la importación de conjuntos de datos preexistentes que albergaban información pertinente a cursos, horarios de clases, capacidad de aulas y la demanda estudiantil.

Figura 7
Cargado de datos.

```
# Carga de datos
#df = pd.read_csv('../data/period_sem_table.csv')
dfBeta = pd.read_csv('../data/period_sem_table_modified.csv')
df = dfBeta.round()
```

Con el fin de obtener un conocimiento más profundo de los datos, se llevó a cabo un análisis exploratorio exhaustivo. Durante este proceso, se identificaron patrones y tendencias significativas que contribuyeron a una comprensión más clara de las necesidades y preferencias de la institución con respecto a la distribución de grupos y horarios.

Figura 8
Exploración de datos



El modelo de aprendizaje automático fue entrenado utilizando la técnica de Random Forest, la cual ha demostrado ser altamente eficaz en la tarea de predecir y recomendar configuraciones de grupos y horarios.

El modelo generó proyecciones basadas en la información de entrada y los resultados se extrajeron en un formato CSV. Estas proyecciones ofrecieron recomendaciones de alto valor para la planificación de grupos estudiantiles.

Figura 9
Predicciones.

```
# Redondea los valores del array al entero más cerca
rounded_arr = np.around(predictions)

# Imprime el array redondeado
#print(rounded_arr)

for i in range(9):
  print(f"Semestre {i+1}: {rounded_arr[i]}")
```

```
[9]
...  Semestre 1: 6.0
     Semestre 2: 22.0
     Semestre 3: 19.0
     Semestre 4: 18.0
     Semestre 5: 15.0
     Semestre 6: 17.0
     Semestre 7: 16.0
     Semestre 8: 15.0
     Semestre 9: 11.0
```

VI. Conclusiones y trabajo futuro

En esta investigación, se aborda el proceso de generación de horarios por semestre para la carrera de Ingeniería en Sistemas Computacionales. Actualmente, este proceso se realiza de manera manual al final de cada semestre con el objetivo de proyectar la cantidad de grupos que se deben habilitar para el siguiente periodo académico.

Aunque hasta ahora el proceso manual ha brindado un pronóstico confiable, requiere una cantidad significativa de horas hombre para llevarlo a cabo.

Para resolver este desafío y mejorar la eficiencia del proceso de generación de horarios, se puede considerar la aplicación de técnicas de aprendizaje automático.

El desarrollo de sistemas de software para resolver este problema siempre es de alto valor. Una herramienta propuesta en esta investigación es generar los horarios de manera automática usando aprendizaje automático, como lo es un algoritmo de aprendizaje supervisado. Como sabemos la matrícula de la carrera de Sistemas Computacionales ha estado en constante crecimiento por tal motivo es necesario que se automatice este proceso.

Como trabajo futuro se llevará a cabo el uso de técnicas de aprendizaje supervisado como lo es el uso de algoritmos de clasificación de bayes ingenuo (Naïve Bayes), para que con esas técnicas se lleve a cabo el aprendizaje y se haga la predicción de manera automática con la información que actualmente nos proporciona el sistema SIE.

En resumen, el análisis del proceso manual existente y la presentación de información relevante permitirían identificar los desafíos y las limitaciones en la generación de horarios. Basándose en esto, se explorarían soluciones de aprendizaje automático y se desarrollaría un sistema automatizado para mejorar la eficiencia y la precisión en la generación de horarios en la institución educativa.

VII. Referencias

[1] J. O. Yunga Pedraza, «Estudio del estado del arte sobre la predicción de deserción universitaria usando machine learning,» de Universidad Salesiana, Ecuador, 2023.

[2] A. U. Castaneda, «Un viaje hacia la inteligencia artificial en la educación,» Realidad y Reflexion, pp. 121-136, 2022.

[3] C. Russo, «Tratamiento Masivo de Datos Utilizando Técnicas de Machine Learning,» REDI, pp. 131-134, 2016.

[4] A. D. Luca, «Uso de la Técnica de Transfer Learning en Machine Learning para la Clasificación de Productos en el Banco Alimentario de La Plata,» SEDICI, pp. 1-16, 2021.

[5] B. A. A. BENITEZ, GENERACION DE HORARIOS MEDIANTE SISTEMAS, México, D.F.: Instituto Politécnico Nacional. Centro de Investigación en Computación, 2007.

[6] E. B. Cañón, «Modelo predictivo del progreso en el aprendizaje de los estudiantes de uniminuto aplicando técnicas de machine learning,» de Scielo, Mexico, 2021.

[7] D. Hinestroza Ramírez, «El Machine Learning a través de los tiempos, y los aportes a la humanidad,» Universidad Libre, pp. 1-17, 2019.

[8] keepcoding, «keepcoding.io,» 2 diciembre 2022. [En línea]. Available: https://keepcoding.io/blog/ciclo-de-vida-de-un-proyecto-en-machine-learning/.

[9] O. Simeone, «A Very Brief Introduction to Machine Learning With Applications to Communication Systems,» de IEEE, 2018.

[10] B. Zarco García, «Algoritmos de clasificación supervisados y semi-supervisados: análisis y comparativa,» UPM, pp. 1-12, 2020.

[11] "Introduction to Machine Learning with Python" de Andreas C. Müller y Sarah Guido. Libro "Pattern Recognition and Machine Learning" de Christopher M. Bishop.

[12] "Introduction to Information Retrieval" de Christopher D. Manning, Prabhakar Raghavan, y Hinrich Schütze.

[13] "An Introduction to the Analysis of Variance" de Ronald A. Fisher.

[14] "Logistic Regression: A Self-Learning Text" de David G. Kleinbaum y Mitchel Klein. Libro "Introduction to the Practice of Statistics" de David S. Moore, George P. McCabe y Bruce A. Craig.

[15] "The Elements of Statistical Learning" de Trevor Hastie, Robert Tibshirani y Jerome Friedman.

[16] “Support Vector Machines" de Nello

Cristianini y John Shawe-Taylor.
[17] Víctor Fabio Suarez, Omar Danilo Castrillón, «Diseño de una metodología basada en técnicas inteligentes para la distribución de procesos Académicos en ambientes de trabajo job shop.
[18] Michael W. Carter, "A Comprehensive Course Timetabling and Student Scheduling System at the University of Waterloo" 2001
[19] Elias Ventura, Eva Marcela, Mendoza Pacas, Carlos Rafael, "Análisis y diseño de un planificador automatizado de horarios universitarios", 2002
[20] Mireya Flores Pichardo, "Revisión de Algoritmos Genéticos Aplicados al Problema de la Programación de Cursos Universitarios" 2011.