UIDE
Powered by
Arizona State University

Maestría en
SISTEMAS DE INFORMACIÓN

Mención **Inteligencia de Negocios y Analítica de Datos Masivos.**

Tesis previa a la obtención del título de Magíster en Sistemas de Información mención Inteligencia de Negocios y Analítica de Datos Masivos.

AUTORES: Ing. Segura Torres Mixi Joselyne
Lcdo. Navas Espin, William Ricardo
Ing. Parra López, Rodolfo Alfredo
Lcdo. Rivera Velástegui, Francisco Nicolás

TUTOR: Ing. Vizcaino Imacaña Fernanda Paulina, PhD.

Diseño de un modelo predictivo de aprobación de materias mediante la analítica de datos para la planificación y proyección de cursos por periodo académico en la Carrera de Telemática - Universidad de Guayaquil.

APROBACIÓN DEL TUTOR

Yo, Paulina Vizcaíno, certifico que conozco los autores/as del presente trabajo siendo los responsables exclusivos tanto de su originalidad y autenticidad, como de su contenido.

Ing. Paulina Vizcaíno Ed. D

DIRECTORA DE TESIS

CERTIFICACIÓN DE AUTORÍA

Nosotros, **SEGURA TORRES MIXI JOSELYNE, NAVAS ESPIN WILLIAM RICARDO, PARRA LOPEZ RODOLFO ALFREDO Y RIVERA VELÁSTEGUI FRANCISCO NICOLÁS,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador, para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Mixi Segura Torres

SEGURA TORRES MIXI JOSELYNE
C.I.: 0952913457

Firmado electrónicamente por:
WILLIAM RICARDO
NAVAS ESPIN

NAVAS ESPIN WILLIAM RICARDO
C.I.: 0918326075

Firmado electrónicamente por:
RODOLFO ANTONIO
PARRA LOPEZ

PARRA LOPEZ RODOLFO ALFREDO
C.I.:0909770448

Francisco Rivera Velastegui

RIVERA VELÁSTEGUI FRANCISCO NICOLÁS
C.I.: 1718171075

Dedicatorias y Agradecimientos

Nosotros, Segura Torres Mixi Joselyne, Navas Espin William Ricardo, Parra López Rodolfo y Rivera Velástegui Francisco Nicolás, expresamos nuestra dedicatoria y más grande agradecimiento a Dios por todas las oportunidades presentes en nuestras vidas, a nuestros familiares; cuyo respaldo inquebrantable, amor y apoyo han sido pilares fundamentales en esta etapa de desarrollo profesional. Este logro que hoy celebramos no sería posible sin la influencia positiva y el apoyo constante de toda nuestra familia y eterno. ¡Gracias!

Agradecemos a la Universidad de Guayaquil, Facultad de Ingeniería Industrial, Carrera Ingeniería en Telemática, por permitirnos recabar la información necesaria para poder realizar la presente investigación, con el fin de que el resultado propuesto, en un futuro sea de mucha utilidad para los procesos de planificación interna de la institución.

ÍNDICE GENERAL

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

En la gestión de instituciones académicas, la planificación es fundamental para asegurar la eficiencia y el logro de objetivos educativos. Esto se traduce en decisiones cruciales sobre la organización de recursos humanos y la estructuración de contenidos según la malla curricular. Estas decisiones impactan directamente en la experiencia de aprendizaje de los estudiantes, así como en la calidad y reputación general de la institución educativa. La planificación, por lo tanto, se posiciona como un pilar esencial para el éxito y desarrollo tanto de las instituciones como de los estudiantes.

La planificación académica es el cimiento sólido sobre el cual se construye la calidad educativa. Permite a las instituciones establecer una estructura coherente y progresiva que garantiza que los estudiantes adquieran una comprensión profunda y sólida de la materia, lo que es esencial para su éxito futuro. Una planificación cuidadosa también puede ayudar a la institución a gestionar eficazmente sus recursos, como aulas y personal docente. La simplificación de la gestión administrativa debe ir de la mano de la incorporación de tecnología, mediante la implementación de programas que agilicen los procesos y reduzcan el tiempo requerido (Huamantumba, 2020).

En este trabajo de tesis, exploraremos la relevancia de la planificación en el ámbito académico desde múltiples perspectivas. Analizaremos el proceso actual de toma de decisiones referente a paralelos a aperturar, destacando cómo una planificación enfocada en modelos predictivos, tiene un resultado efectivo, mejorando la experiencia educativa. Asimismo, consideraremos cómo la planificación es crucial para la optimización de los recursos, lo que contribuye a la sostenibilidad y el éxito de las instituciones académicas en un entorno evolutivo. En última instancia, esta investigación resaltará la importancia de la planificación como un proceso estratégico que contribuye en gran medida al logro de la reducción de costos innecesarios y al fortalecimiento de la calidad en la educación superior.

PALABRAS CLAVES

Modelo predictivo, Python, Cassandra, base de datos, planificación académica, universidad.

ABSTRACT

In the management of academic institutions, planning is essential to ensure efficiency and the achievement of educational objectives. This translates into crucial decisions regarding the organization of human resources and the structuring of content according to the curriculum. These decisions directly impact students, learning experience, as well as the overall quality and reputation of the educational institution. Planning, therefore, is positioned as a fundamental pillar for the success and development of both institutions and students.

Academic planning is the solid foundation on which educational quality is built. It allows institutions to establish a coherent and progressive structure that ensures students acquire a deep and solid understanding of the subject, which is essential for their future success. Careful planning can also help the institution effectively manage its resources, such as classrooms and teaching staff. Streamlining administrative management should go hand with hand, with the incorporation of technology through the implementation of programs that streamline processes and reduce the required time (Huamantumba, 2020).

In this thesis, we will explore the relevance of planning in the academic field from multiple perspectives. We will analyze the current decision-making process regarding the opening of parallel courses, highlighting how planning focused on predictive models has an effective outcome, improving the educational experience. Additionally, we will consider how planning is crucial for optimizing resources, contributing to the sustainability and success of academic institutions in an evolving environment. Ultimately, this research will emphasize the importance of planning as a strategic process that significantly contributes to achieving the reduction of unnecessary costs and strengthening quality in higher education.

KEYWORDS

Predictive model, Python, Cassandra, database, academic planning, university.

CAPITULO I

1.1. INTRODUCCIÓN

En el contexto laboral y académico, la eficiencia y la efectividad en la planificación de cursos son fundamentales para garantizar una oferta académica de calidad. En este sentido, el presente trabajo de tesis aborda un problema crítico que afecta la toma de decisiones en el ámbito educativo: la falta de información oportuna en el proceso de planificación académica. Esta carencia de información precisa conduce a dificultades y errores en la toma de decisiones sobre el número de cursos a abrir en cada nivel académico, lo que, a su vez, impacta en la distribución de recursos y en la asignación de docentes. La presente introducción busca contextualizar el problema, resaltando su relevancia y justificando la necesidad de abordarlo.

En la actualidad, la planificación académica en instituciones educativas depende en gran medida de la carga de notas finales subidas al sistema. Si bien esta práctica ha sido valiosa durante años, se ha vuelto insuficiente y limitante en un entorno académico caracterizado por la creciente demanda de información precisa y oportuna. La dependencia de las notas finales como único indicador para la planificación académica plantea desafíos significativos en términos de eficiencia y eficacia, lo que afecta directamente la calidad de la oferta académica (Reyes-González, 2022).

El problema central radica en la falta de información actualizada y detallada que permita a los responsables de la planificación académica tomar decisiones informadas y estratégicas. La carencia de datos precisos dificulta la evaluación de la demanda de cursos, la asignación de recursos, la optimización de horarios y la distribución equitativa de cargas académicas entre docentes (Gupta, 2020). Como resultado, se generan dificultades innecesarias, malentendidos y la necesidad de reaccionar a situaciones imprevistas, lo que impacta en la calidad general de la experiencia educativa.

El proceso de planificación académica es una tarea compleja que requiere una visión estratégica y una gestión eficiente. La toma de decisiones fundamentales, como la apertura de cursos, la asignación de docentes y la distribución de recursos, debe basarse en datos fiables y actualizados. La planificación académica efectiva no solo garantiza que los recursos disponibles se utilicen de manera óptima, sino que también permite anticipar y abordar desafíos futuros.

Este trabajo de tesis se centra en la importancia de contar con información precisa y oportuna para la toma de decisiones en la planificación académica. Aborda el problema existente en la toma de decisiones en el ámbito laboral y académico, donde la falta de información adecuada ha llevado a dificultades y errores en la planificación de cursos. El problema no solo impacta en la distribución de recursos y en la asignación de docentes, sino que también afecta la calidad de la oferta académica en su conjunto.

Para abordar este problema, es esencial entender su alcance y dimensiones. En este sentido, se analizarán las limitaciones actuales de la planificación académica, identificando las áreas críticas que requieren mejoras. Además, se explorarán las soluciones tecnológicas y metodológicas disponibles para obtener información precisa y oportuna en el proceso de planificación académica. La tesis tiene como objetivo proponer una solución efectiva que permita a las instituciones educativas superar estos obstáculos y optimizar la gestión de recursos y docentes.

1.2. PREGUNTA DE INVESTIGACIÓN

¿Cómo optimizar el proceso de planificación académica utilizando la analítica de datos de 10 periodos académicos en la carrera Telemática Universidad de Guayaquil?

1.3. OBJETIVO GENERAL

Diseñar un modelo predictivo de aprobación de materias mediante la analítica de datos para la planificación y proyección de cursos por periodo académico en la carrera de Telemática Universidad de Guayaquil.

1.4. OBJETIVOS ESPECÍFICOS

1.- Investigar diferentes algoritmos y enfoques de aprendizaje automático, como la clasificación binaria y los modelos de regresión, para determinar cuál es el más adecuado para este contexto específico.

2.- Aplicar técnicas de analítica de datos para identificar patrones y tendencias en los datos académicos de los estudiantes de la carrera de Telemática.

3.- Desarrollar un modelo predictivo utilizando técnicas de aprendizaje automático (machine learning) que sea capaz de predecir la aprobación de materias para mejorar los resultados de la planificación en función de las variables seleccionadas.

1.5. DEFINICIÓN DEL PROBLEMA

El problema que se presenta en el ámbito laboral es la falta de información oportuna en el proceso de planificación académica, lo que conlleva a dificultades y errores en la toma de decisiones sobre el número de cursos a abrir en cada nivel académico. Actualmente, se depende de la carga de notas finales subidas al sistema para realizar esta planificación, lo que limita el tiempo disponible y genera un dilema en la toma de decisiones. Esta falta de información afecta la eficiencia y efectividad en la planificación de cursos, lo cual puede tener consecuencias negativas en la distribución de recursos, asignación de docentes y la calidad de la oferta académica (Auccapuri, 2021). Por lo tanto, es necesario abordar este problema y buscar soluciones que permitan contar con información precisa y oportuna para la toma de decisiones en la planificación académica.

CAPITULO II

2.1. METODOLOGÍA

La metodología propuesta a implementar en el proceso de la presente investigación será la procedimental, la cual permitirá desarrollar un modelo predictivo robusto y confiable que pueda ser utilizado para mejorar la planificación académica y la proyección de cursos en la carrera de Telemática, facilitando la toma de decisiones informadas y eficientes.

La metodología procedimental consta de los siguientes pasos:

- Definición del problema
- Recopilación de datos
- Análisis exploratorio de datos
- Preparación y limpieza de datos
- Selección de variables y características
- Diseño del modelo predictivo
- Análisis del modelo
- Interpretación de resultados

2.2. NECESIDADES, STAKEHOLDERS Y ÁREAS DEL NEGOCIO INVOLUCRADAS

Actualmente en la Facultad de Ingeniería Industrial es necesario el diseño de un modelo predictivo de aprobación de materias para los procesos semestrales como la planificación académica y la apertura de nuevos cursos en la carrera de Ing. Telemática ya que enfrenta dificultades y errores debido a la falta de información oportuna.

La insuficiencia de datos representa un obstáculo significativo que obstaculiza la capacidad de tomar decisiones precisas en lo que respecta a la determinación del número de cursos que se deben abrir en cada nivel académico. Como resultado de esta limitación, la planificación se ve forzada a depender de la carga de notas finales que se haya registrado en el sistema, lo que no solo restringe el margen de tiempo disponible para llevar a cabo esta tarea, sino que también da lugar a un dilema en el proceso de toma de decisiones.

La carencia de información completa y oportuna tiene un impacto directo en la eficiencia y efectividad de la planificación semestral de cursos. Esto, a su vez, puede generar repercusiones negativas en varios aspectos fundamentales, incluyendo la distribución de recursos, la asignación de docentes y, en última instancia, la calidad de la oferta académica. En este sentido, es necesario abordar este problema de manera sistemática y buscar soluciones que permitan recopilar información precisa y actualizada, lo que, a su vez, respaldará una toma de decisiones más informada y estratégica en el proceso de planificación académica.

Las áreas involucradas en la planificación académica semestral:

1. Dirección de carrera
2. Coordinación académica
3. Gestoría pedagógica curricular
4. Docentes académicos
5. Departamento TI
6. Alumnado
7. Gestión de Recursos Humanos

2.3. ALCANCE DEL PROYECTO

- Mediante la data de los estudiantes de los 10 periodos académicos anteriores al analizado para la predicción, se realizará un modelo de recomendación de apertura de cursos en la carrera de Telemática.
- La investigación se basará en la recopilación, depuración y análisis de datos históricos de los estudiantes de la carrera de Telemática, como sus calificaciones en cursos anteriores, asistencia, participación en actividades académicas, entre otras variables relevantes.
- La información obtenida se utilizará para planificar y proyectar los cursos que se ofrecerán en el periodo analizado, con el objetivo de optimizar la apertura de los cursos.
- Por medio del modelo predictivo escogido, se podrá visualizar mediante los resultados obtenidos la sugerencia de cursos que se ofrecerán en el periodo académico analizado.
- Se espera mediante el modelo predictivo, obtener un grado de confianza de estos resultados, esto se verá reflejado en el porcentaje de acierto de la proyección inicial versus los resultados del modelo predictivo. Esto será controlado por los KPI's propuestos.

2.4. IMPACTO DE NEGOCIO

El presente proyecto estará enfocado en la mejora y la innovación de un proceso que actualmente se lo realiza de forma manual con escasez de información para tomar decisiones, lo que produce una pérdida de tiempo al tener rehacer el proceso hasta tenerlo viable según las necesidades reales (Vargas-Larraguível, 2021). El problema actual ha producido que se tenga que cerrar paralelos al no haber completado el número mínimo de estudiantes necesarios. Por medio del modelo predictivo escogido, se podrá visualizar mediante los resultados obtenidos la sugerencia de cursos que se ofrecerán en el periodo académico analizado.

Adicionalmente es relevante destacar que la optimización que hemos mencionado tendrá un efecto doblemente beneficioso en términos de reducción de costos y tiempo empleado. Al minimizar los tiempos requeridos y mejorar el grado de ajuste en la optimización de recursos

(Balzer, 2020), no solo lograremos una eficiencia operativa mejorada, sino que también generaremos un impacto financiero positivo de valor agregado.

La reducción de tiempos en los procesos internos conlleva una disminución directa de los costos asociados con los docentes inmersos en el proceso de planificación, los recursos materiales y otros gastos operativos. Esto se traduce en ahorros tangibles que pueden destinarse a otros aspectos cruciales como la inversión en investigación y desarrollo para la facultad (Panchi Arias, 2021).

2.5. CRONOGRAMA Y PLANIFICACIÓN

Nuestra planificación inicia en el mes de agosto con la materia de Plan de Titulación y sistema integrados de gestión, donde definimos nuestro tema y los objetivos alcanzar en nuestro proyecto de tesis. También se recoleto los datos de los últimos 10 periodos académicos de la carrera de Telemática, se realizó la literatura correspondiente apoyado con la materia de almacenamiento e integración de datos, la cual se eligió base de datos Cassandra.

Figura 1:

Cronograma de la tesis

	1				2				3			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Aprobación el protocolo y designación de director												
Revisión de literaruta												
Desarrollo del capítulo 1												
Revisión de literaruta												
Desarrollo del capítulo 2												
Recolección de información												
Entrevistas a directivos de la Carrera												
Revisión de literaruta												
Realización de gráficas y modelos												
Procesamiento de datos												
Desarrollo del capítulo 3												
Análisis de resultados												
Conclusiones y recomendaciones												
Presentación de informe preliminar												
Presentación de informe final												

2.6. DESARROLLO

2.6.1. DEFINIR LAS FASES DENTRO DEL PROYECTO

Definición de Objetivos y Alcance:

- Identificar claramente los objetivos del modelo predictivo, como predecir el éxito o el fracaso en la aprobación de materias.
- Determinar el alcance del proyecto, incluyendo los cursos, períodos académicos y otros factores relevantes.

Recopilación de Datos:

- Reunir datos históricos relacionados con la carrera de Telemática, como registros académicos de estudiantes, información de cursos, notas, fechas importantes, etc.
- Limpiar y preprocesar los datos para eliminar valores atípicos y asegurar su calidad.

Análisis Exploratorio de Datos (EDA):

- Realizar un análisis descriptivo de los datos para identificar patrones, tendencias y relaciones.
- Visualizar los datos para comprender mejor su distribución y características.

Selección de Características:

- Identificar las variables que más influyen en el rendimiento académico de los estudiantes y seleccionar las más relevantes para el modelo.

Preparación de Datos:

- Dividir los datos en conjuntos de entrenamiento, validación y prueba.
- Normalizar o estandarizar las características si es necesario.
- Realizar codificación de variables categóricas.

Desarrollo del Modelo Predictivo:

- Seleccione el algoritmo de aprendizaje automático adecuado para el problema de predicción de aprobación de materias.

- Entrenar el modelo utilizando el conjunto de datos de entrenamiento.
- Ajustar hiperparámetros y evaluar su rendimiento en el conjunto de validación.

Evaluación del Modelo:

- Utilizar métricas de evaluación adecuadas, como precisión, recuperación, puntuación F1, entre otras, para medir el rendimiento del modelo.
- Realizar validación cruzada para evaluar la robustez del modelo.

Ajuste y Optimización del Modelo:

- Realice ajustes en el modelo, como cambiar algoritmos, modificar características o ajustar hiperparámetros, según sea necesario para mejorar su rendimiento.

Interpretación de resultados

- Se interpretará el modelo para comprender qué variables y características tienen un mayor impacto en las predicciones. Esto proporcionará información valiosa sobre los factores que influyen en la aprobación de materias y puede utilizarse para tomar decisiones informadas en la planificación de cursos.

2.6.2. DEFINIR LAS FUENTES DE INFORMACIÓN

Fuentes Internas:

- **Registros Académicos de la UG. -** La propia universidad es una fuente interna importante que proporciona datos sobre la matrícula de estudiantes, notas de materias, fechas de inscripción, y más.
- **Gestoría de planificación académica de Telemática. -** El departamento de Telemática de la UG puede proporcionar datos específicos relacionados con su programa de estudios, cursos ofrecidos y requisitos académicos.
- **Entrevistas Internas. -** Los resultados de encuestas y evaluaciones internas realizadas por la universidad pueden contener información valiosa sobre la satisfacción de los estudiantes y su rendimiento académico.
- Reglamentos internos vigentes de la Universidad de Guayaquil en base a los datos requeridos.

Fuentes Externas:

- **CES consejo de educación superior. -** Su función es ser el organismo planificador, regulador y coordinador del Sistema Nacional de Educación Superior de la República del Ecuador.

- **CACES. -** Es un organismo público que tiene la responsabilidad de regular, planificar y coordinar el Sistema de Aseguramiento de la Calidad de la Educación Superior en Ecuador.

- **Fuentes de Investigación Académica. -** La literatura académica sobre la educación superior, la planificación de cursos y la retención estudiantil puede proporcionar información valiosa.

2.6.3. ANÁLISIS DE MOTORES DE BASE DE DATOS

Realizaremos la construcción de un modelo de recomendación de apertura de cursos para la carrera de Telemática en la Universidad De Guayaquil, basado en la información de los estudiantes de los 10 períodos académicos anteriores al que está siendo analizado. El proceso investigativo se sustentará en la recopilación, depuración y análisis de datos históricos de los estudiantes, contemplando aspectos como sus calificaciones previas, asistencia, participación en actividades académicas y otras variables significativas que se seleccionarán en el proceso investigativo.

La concepción de un modelo predictivo para la aprobación de materias mediante el empleo de técnicas analíticas y la exploración de datos constituye un paso fundamental en este proceso. Esto permitirá una mejor planificación y proyección de los cursos para cada periodo académico en la carrera de Telemática en la Universidad de Guayaquil. Los objetivos específicos son los siguientes:

- Se llevará a cabo una investigación exhaustiva de diversos algoritmos y enfoques de aprendizaje automático. Esto incluirá la evaluación de métodos como la clasificación binaria y los modelos de regresión. El objetivo es determinar cuál de estos enfoques es el más apropiado para la tarea específica de predicción en este contexto.
- Serán aplicadas técnicas de analítica de datos con el propósito de identificar patrones y tendencias en los registros académicos de los estudiantes de Telemática. Este análisis permitirá descubrir relaciones ocultas y comprender el comportamiento académico de los estudiantes a lo largo de los períodos anteriores.
- Utilizando técnicas avanzadas de aprendizaje automático, se creará un modelo predictivo. Este modelo estará diseñado para anticipar la aprobación de materias con base en las variables seleccionadas previamente. El objetivo es proporcionar una herramienta confiable que mejore la precisión en la planificación de cursos y optimice los resultados académicos.

La investigación integral que involucra la recopilación y análisis de datos históricos de estudiantes de Telemática serán la base para el diseño y desarrollo de un modelo predictivo que permitirá mejorar la planificación y proyección. Este enfoque, respaldado por técnicas de aprendizaje automático y analítica de datos, busca optimizar la experiencia educativa en la carrera (Contreras, 2020).

2.6.3.1. MONGODB

MongoDB es una base de datos NoSQL de código abierto que se caracteriza por su enfoque en el almacenamiento y la gestión de datos en formato de documentos. A diferencia de las bases de datos relacionales tradicionales que utilizan tablas y filas, MongoDB almacena los datos en documentos BSON (Binary JSON) (Narváez, 2020), lo que permite una mayor flexibilidad y escalabilidad en el modelado de datos.

Figura 2:

Logo de MongoDB.

Nota: Obtenido de la web. Elaborado por MongoDB, Inc.

CARACTERÍSTICAS PRINCIPALES

MongoDB almacena los datos en documentos que pueden contener campos y valores de diferentes tipos, incluyendo cadenas, números, fechas, matrices y documentos anidados, además, no requiere un esquema fijo y predefinido, esto permite una mayor agilidad en el desarrollo, ya que los cambios en la estructura de los datos no requieren modificaciones complejas en el esquema de la base de datos (Narváez, 2020).

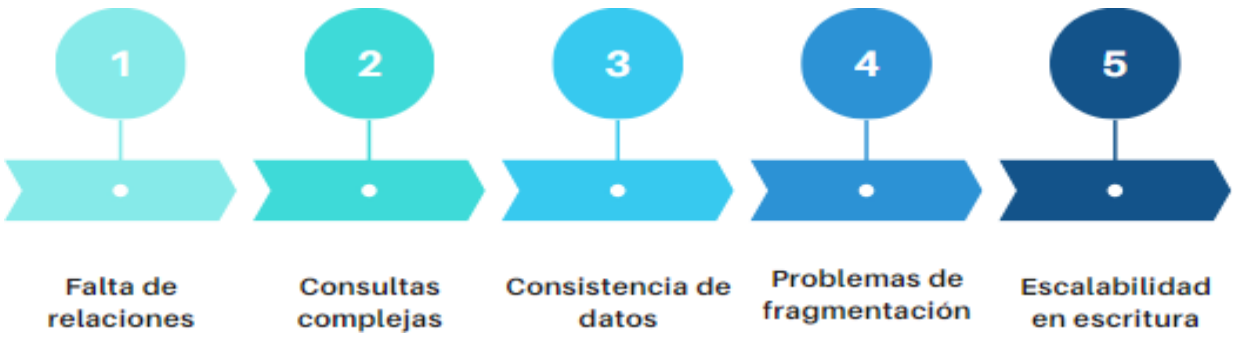
MongoDB está diseñado para escalar horizontalmente, lo que significa que puedes distribuir tus datos en Múltiples servidores para manejar cargas de trabajo crecientes, también almacena datos en formato de documentos, tiene la capacidad de persistir esos datos en disco, lo que garantiza la durabilidad de los datos incluso después de un reinicio del sistema. MongoDB admite consultas complejas y búsqueda de datos mediante su lenguaje de consulta basado en JSON,

también admite la replicación para garantizar la disponibilidad y la redundancia de los datos y proporciona un equilibrio de carga automático para distribuir las operaciones de lectura y escritura entre los nodos del clúster (Chauhan, 2019).

DESVENTAJAS

Figura 3:

Desventajas de MongoDB



1. **Falta de relaciones**
   A diferencia de las bases de datos relacionales, MongoDB no maneja naturalmente las relaciones entre datos. Si tu aplicación requiere relaciones complejas entre entidades, modelarlas en MongoDB puede resultar complicado y requerir estructuras de datos adicionales (Chango Gavilánez, 2016).

2. **Consultas complejas**
   Aunque MongoDB ofrece un lenguaje de consulta flexible, las consultas complejas que involucran varias colecciones y relaciones pueden ser más difíciles de escribir y optimizar en comparación con las bases de datos relacionales.

3. **Consistencia de datos**
   MongoDB ofrece niveles de consistencia configurables, pero en entornos de alta disponibilidad y particiones de red, puede haber casos en los que se comprometa la consistencia de los datos entre los nodos del clúster.

4. **Problemas de fragmentación**
   A medida que se insertan, actualizan y eliminan documentos, MongoDB puede sufrir de fragmentación de datos, lo que puede afectar el rendimiento y la eficiencia del almacenamiento.

5. **Escalabilidad en escritura**
   Aunque MongoDB es escalable horizontalmente, la escalabilidad en escritura puede ser desafiante en ciertos casos de carga de trabajo intensivo de escritura debido a la estructura de documentos y las operaciones de indexación.

2.6.3.2. CASSANDRA

Apache Cassandra es una base de datos distribuida y altamente escalable que se diseñó para manejar grandes volúmenes de datos distribuidos a través de múltiples servidores y ubicaciones geográficas. Está diseñado para ofrecer alta disponibilidad, tolerancia a fallos y rendimiento escalable, especialmente en entornos donde se requiere una latencia baja y una alta velocidad de escritura y lectura (Piccardi, 2021).

Figura 4:

Logo de Cassandra.

Nota: Obtenido de la web. Elaborado por Cassandra Inc.

CARACTERÍSTICAS

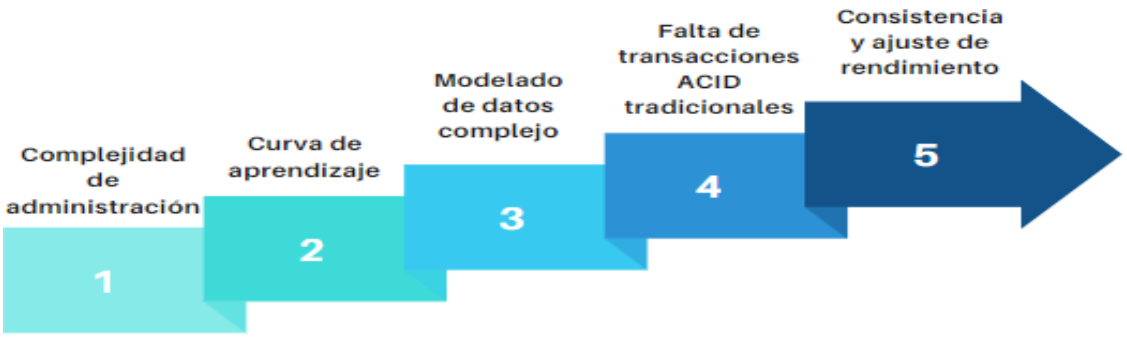
Cassandra utiliza un modelo de arquitectura distribuida, donde los datos se distribuyen en varios nodos en un clúster. Cada nodo es igual en términos de roles y capacidades, lo que elimina los puntos únicos de fallo, su capacidad es de escalar horizontalmente, lo que significa que se pueden agregar nuevos nodos al clúster para manejar cargas de trabajo crecientes, también ofrece replicación automática de datos a través de nodos, lo que garantiza la alta disponibilidad y la tolerancia a fallos (Wahid, 2019).

Cassandra utiliza un modelo de datos tipo columna que permite una gran flexibilidad en la representación y el almacenamiento de datos, ofrece soporte nativo para la replicación de datos entre diferentes centros de datos, lo que permite mantener datos disponibles y consistentes en ubicaciones geográficas distantes, además permite ajustar el nivel de consistencia de lectura y escritura en función de los requisitos de la aplicación, lo que permite equilibrar la coherencia y la disponibilidad.

DESVENTAJAS

Figura 5:

Desventajas de Cassandra



1. **Complejidad de administración**
   Cassandra es una base de datos distribuida y altamente escalable, lo que puede aumentar la complejidad de su administración en comparación con bases de datos más simples. Requiere una comprensión sólida de su arquitectura y configuración para implementar y mantener correctamente un clúster.

2. **Curva de aprendizaje**
   Debido a su naturaleza distribuida y sus características específicas, Cassandra puede tener una curva de aprendizaje empinada para los desarrolladores y administradores que no están familiarizados con bases de datos NoSQL y conceptos de distribución.

3. **Modelado de datos complejo**
   Aunque Cassandra ofrece flexibilidad en el modelado de datos, puede ser más desafiante de modelar correctamente en comparación con bases de datos relacionales tradicionales. El modelado incorrecto puede llevar a ineficiencias en las consultas y el rendimiento.

4. **Falta de transacciones ACID tradicionales**
   Aunque Cassandra ofrece transacciones y coherencia, no sigue el modelo de transacciones ACID (Atomicidad, Consistencia, Aislamiento, Durabilidad) que ofrecen muchas bases de datos relacionales.

5. **Consistencia y ajuste de rendimiento**
   Configurar la consistencia adecuada para las operaciones de lectura y escritura puede ser un desafío. Además, ajustar el rendimiento de Cassandra para que se adapte a la carga de trabajo específica puede requerir pruebas y ajustes exhaustivos.

2.6.3.3. NEO4J

Neo4j es una base de datos orientada a gráficos, lo que significa que está diseñada para almacenar y gestionar datos en forma de nodos interconectados por relaciones (Link, 2020). A diferencia de las bases de datos relacionales tradicionales que utilizan tablas y filas para almacenar datos, las bases de datos orientadas a gráficos se basan en el concepto de nodos y relaciones, lo que las hace ideales para modelar y consultar datos altamente relacionales y complejos.

Figura 6:

Logo de *Neo4j.*

Nota: Obtenido de la web. Elaborado por Neo4j, Inc.

PRINCIPALES CARACTERÍSTICAS

Neo4J representa una base de datos de grafos de excelente desempeño creada en el lenguaje de programación Java. Sus características fundamentales comprenden el almacenamiento, la administración y la interrogación de datos altamente relacionados con una eficiencia sobresaliente. Estos datos se encuentran representados en un modelo de grafo en el cual se almacenan como nodos interconectados mediante relaciones, lo que posibilita la construcción y exploración eficaz de relaciones complejas. Neo4j emplea un lenguaje de consulta denominado Cypher, diseñado específicamente para interactuar con datos de grafo. Cypher habilita la ejecución de consultas enfocadas en patrones de relaciones y travesías de grafo. Esta herramienta es especialmente indicada para situaciones en las que la comprensión de las interconexiones entre los datos resulta fundamental (Ayinuer, 2022).

DESVENTAJAS

Figura 7:

Desventajas de Neo4j

1. **Complejidad**
   Aunque Neo4j es una herramienta poderosa, su modelo de datos basado en gráficos puede ser más complejo de entender y modelar en comparación con las bases de datos relacionales tradicionales. Esto puede requerir un tiempo de aprendizaje adicional para los desarrolladores.

2. **Escalabilidad horizontal limitada**
   Si bien Neo4j ha mejorado en términos de escalabilidad horizontal en comparación con versiones anteriores, todavía puede ser menos eficiente en términos de escalabilidad en comparación con algunas otras tecnologías de bases de datos NoSQL.

3. **Tamaño de almacenamiento**
   Neo4j requiere una cantidad significativa de memoria para funcionar eficientemente, especialmente cuando se manejan conjuntos de datos grandes y altamente conectados. Esto puede limitar su capacidad para ejecutarse en sistemas con recursos limitados.

4. **Rendimiento en ciertas consultas**
   Aunque Neo4j está optimizado para consultas de gráficos, no es la mejor opción para todas las consultas y tipos de datos. Algunas operaciones específicas pueden ser más lentas o menos eficientes en comparación con otras bases de datos especializados.

5. **Integración con otras tecnologías**
   Aunque Neo4j tiene integración con varios lenguajes de programación y frameworks populares, es posible que no esté tan bien integrado en ciertas tecnologías o ecosistemas como las bases de datos relacionales.

2.6.3.4. REDIS

Redis es una base de datos en memoria de código abierto que se utiliza como almacén de datos en caché, almacén de datos en tiempo real y almacén de estructuras de datos. Se destaca por su alta velocidad y rendimiento, ya que almacena datos en la memoria principal en lugar de en la discoteca, lo que permite tiempos de acceso extremadamente rápidos.

Figura 8:

Logo de *Redis.*

Nota: Obtenido de la web. Elaborado por Redis, Inc.

PRINCIPALES CARACTERÍSTICAS

Redis representa una base de datos distribuida de tipo clave-valor, perteneciente a la categoría NoSQL, que opera en memoria y es de código abierto. Su particularidad radica en su capacidad para admitir múltiples estructuras de datos y transacciones, así como su notoriedad surge de su

destacada velocidad y baja latencia, cualidades especialmente valiosas en escenarios donde la rapidez de acceso a los datos es fundamental gracias a su lenguaje. Sus aplicaciones más comunes incluyen servir como base de datos, desempeñar funciones de caché y operar como intermediario en sistemas de mensajería (Abu Kausar, 1532-1540). Al ser Redis una base de datos que opera en memoria proporciona una velocidad elevada y una diversidad de funcionalidades para atender variadas exigencias en el ámbito del almacenamiento y recuperación de datos. En esta solución, todos los datos se almacenan en la memoria primaria, lo cual habilita un acceso sumamente veloz a la información. Esta característica la convierte en la elección óptima para aplicaciones que demandan respuestas inmediatas.

DESVENTAJAS

Figura 9:

Desventajas de Redis

1. **Capacidad limitada por la memoria**
   Dado que Redis almacena datos en la memoria principal, la cantidad de datos que puede manejar está limitada por la cantidad de RAM disponible en el sistema. Esto puede ser una limitación en comparación con bases de datos que almacenan datos en disco y pueden manejar conjuntos de datos más grandes.

2. **Riesgo de pérdida de datos**
   Aunque Redis ofrece opciones de persistencia en disco, como instantáneas y registros de transacciones, estas no ofrecen la misma garantía de durabilidad que las bases de datos que almacenan datos directamente en disco. En situaciones extremas, como un fallo de energía repentino, es posible que se pierdan datos.

3. **Complejidad de configuración**
   La configuración y el ajuste de Redis pueden ser complicados, especialmente al trabajar con características avanzadas como la replicación, la alta disponibilidad y la clusterización. Una configuración incorrecta puede provocar problemas de rendimiento o pérdida de datos.

4. **Escalabilidad vertical**
   Aunque Redis ofrece clusterización y replicación maestro-esclavo para mejorar la escalabilidad, su escalabilidad vertical (agregar más recursos a un único nodo) puede ser limitada en comparación con otras bases de datos distribuidas.

5. **Consistencia eventual**
   En el modo de replicación, Redis garantiza la consistencia eventual entre los nodos maestros y esclavos, lo que significa que podría haber un pequeño retraso entre las actualizaciones en el nodo maestro y la replicación en los nodos esclavos.

2.6.4. COMPARATIVA ENTRE LAS DIFERENTES BASES DE DATOS

Figura 10:

Comparativa de las Bases de datos

ASPECTOS	neo4j	redis	CASSANDRA	mongoDB.
TIPOS DE DATOS	GRÁFICOS	KEY - VALUE		
MODELO	NOSQL	NOSQL	COLUMN FAMILIES	BSON
ESCALABILIDAD	BUENA	BUENA	CLÚSTERES	CLÚSTERES DE RÉPLICAS
CONSULTAS	CYPHER	COMANDOS	CQL	FLEXIBLES
RELACIONES	SI	NO	NO	NO
FLEXIBILIDAD	ALTA	ALTA	ALTA	ALTA
RENDIMIENTO	VARIABLE	MUY RÁPIDO	MUY RÁPIDO	RÁPIDO
USO PRINCIPAL	GRAFOS	CACHE	IoT	APLICACIONES WEB - MÓVILES

2.6.5. JUSTIFICACIÓN DE USO DE TIPO DE BASE DE DATOS

El proyecto se centra en la adquisición y tratamiento de datos provenientes de un sistema transaccional. Estos datos, inicialmente presentes en archivos de Microsoft Excel, se canalizarán hacia Apache Cassandra, una base de datos orientada a documentos, donde se ejecutará un proceso de análisis y depuración. La etapa de procesamiento implica la extracción selectiva de datos de diversas tablas de información obtenidas directamente desde el sistema de la Universidad de Guayaquil. El enfoque se dirigirá a identificar y extraer campos específicos requeridos para el análisis posterior, la finalidad principal radica en la visualización coherente y comprensible de los datos elegidos. Exponemos algunas formas en las que podríamos interactuar entre las dos bases de datos:

Se importará los datos desde el sistema transaccional a tablas de Excel. Realizamos la importación de las tablas de Excel obtenidas a la base de datos de Apache Cassandra, utilizando bibliotecas diseñadas para facilitar la interacción entre Apache Cassandra y Excel, facilitando la actualización automática de datos en Excel a medida que cambian en Apache Cassandra. Se realizará la limpieza y depuración de la información, seleccionado la data de los 10 períodos anteriores usaríamos Power BI para crear informes o paneles de control empleando datos de Apache Cassandra. La integración de estas dos plataformas de almacenamiento de datos plantea varias posibilidades de interacción (Mihiranga, 2022), se podría establecer un flujo constante de datos, donde los nuevos registros en el sistema transaccional se actualicen automáticamente en Apache Cassandra. Otra opción es la sincronización programada, en la que los datos se actualizan en intervalos predefinidos, manteniendo cierta regularidad en la actualización de la base de datos de Apache Cassandra. Utilizando el conector Cassandra Power BI Connector integraremos los Dashboard en Power BI extrayendo la información desde nuestra data en Apache Cassandra.

En primer lugar, Cassandra es una base de datos de tipo NoSQL, cuyas características intrínsecas la hacen sumamente adecuada para la gestión de datos en gran escala (Anusha, 2021). La naturaleza distribuida de Cassandra permite manejar grandes volúmenes de información generados por estudiantes, cursos, profesores y evaluaciones a lo largo de múltiples periodos académicos, que para objeto del presente proyecto se centrará en los 10 últimos períodos. Esta escalabilidad inherente resulta vital en entornos educativos, donde los datos acumulados pueden crecer exponencialmente con el tiempo.

Por otro lado, la flexibilidad intrínseca de Cassandra le otorga la versatilidad necesaria para adaptarse de manera ágil y dinámica a la evolución constante de los datos. Esta capacidad resulta esencial al desarrollar un modelo predictivo que requiere incorporar y procesar diversos tipos de información, como calificaciones de exámenes, registros de inscripción, antecedentes académicos, notas de exámenes parciales y otros factores influyentes en el resultado del diseño y puesta en marcha del modelo.

Un factor determinante en la elección de Cassandra es su arquitectura distribuida y tolerante a fallos. Esto asegura que los datos estén disponibles y sean accesibles en todo momento, reduciendo el riesgo de interrupciones en el proceso de planificación y proyección de cursos. La posibilidad de replicar y distribuir los datos en varios nodos refuerza la resistencia a fallos, garantizando la integridad y continuidad de la información y sobre todo la disponibilidad de esta cuando se la necesite.

La habilidad de Cassandra para gestionar operaciones de escritura y lectura de manera simultánea y con una notable velocidad adquiere una importancia destacada para la ejecución del presente proyecto (Del Pozo Puñal, 2019). Esta característica se traduce en la agilidad necesaria para llevar a cabo análisis en tiempo real, lo que a su vez habilita la entrega de predicciones y recomendaciones oportunas dirigidas a los responsables de la planificación académica de la Universidad de Guayaquil.

Finalmente, la orientación a columnas de Cassandra se alinea perfectamente con la complejidad de las consultas analíticas requeridas para un modelo predictivo. Facilita el acceso eficiente tanto a los datos históricos como a los datos en tiempo real, esenciales para el modelado predictivo efectivo. La capacidad de Cassandra para gestionar grandes conjuntos de datos sin sacrificar el rendimiento la convierte en una elección sólida y robusta para la creación y despliegue de un modelo predictivo de aprobación de materias en el ámbito académico, permitiendo la planificación y proyección de cursos de manera estratégica y efectiva.

En el dashboard de Power BI se presentará datos relacionados con la educación que podría proporcionar insights significativos sobre el rendimiento académico, los factores que influyen en él y las tendencias a lo largo del tiempo. Este análisis puede ser valioso tanto para tomar decisiones en el ámbito educativo como para identificar áreas de mejora y optimización en el proceso de enseñanza y aprendizaje.

- **Análisis General**
  - Identificación de las facultades y carreras con mayor cantidad de estudiantes y sus respectivos rendimientos.
  - Análisis de la distribución de género y edades en diferentes programas académicos.
  - Evaluación de la tasa de asistencia promedio en los diferentes ciclos lectivos y materias.
  - Exploración de los patrones de aprobación y reprobación de estudiantes en relación con el tipo de nota y las materias.

- **Desempeño por Docente**
  - Comparación del rendimiento de los estudiantes bajo diferentes docentes y su relación con la asistencia.

- Identificación de docentes con altos índices de aprobación y su influencia en el rendimiento académico.

- **Análisis Geográfico**
  - Visualización de la distribución de estudiantes por provincia, cantón y parroquia.
  - Comparación de los resultados académicos entre diferentes áreas geográficas.

- **Desempeño por Ciclo Lectivo**
  - Seguimiento del desempeño de los estudiantes a lo largo de varios ciclos lectivos.
  - Identificación de tendencias de mejora o empeoramiento en los resultados académicos.

- **Estadísticas de Notas**
  - Análisis de la distribución de notas y su relación con el tipo de nota y el estado de aprobación.
  - Identificación de patrones en las notas de los estudiantes a lo largo del tiempo.

- **Estadísticas Demográficas**
  - Desglose del desempeño académico por género y su relación con la edad.
  - Análisis de la relación entre la edad, la provincia de origen y el rendimiento académico.

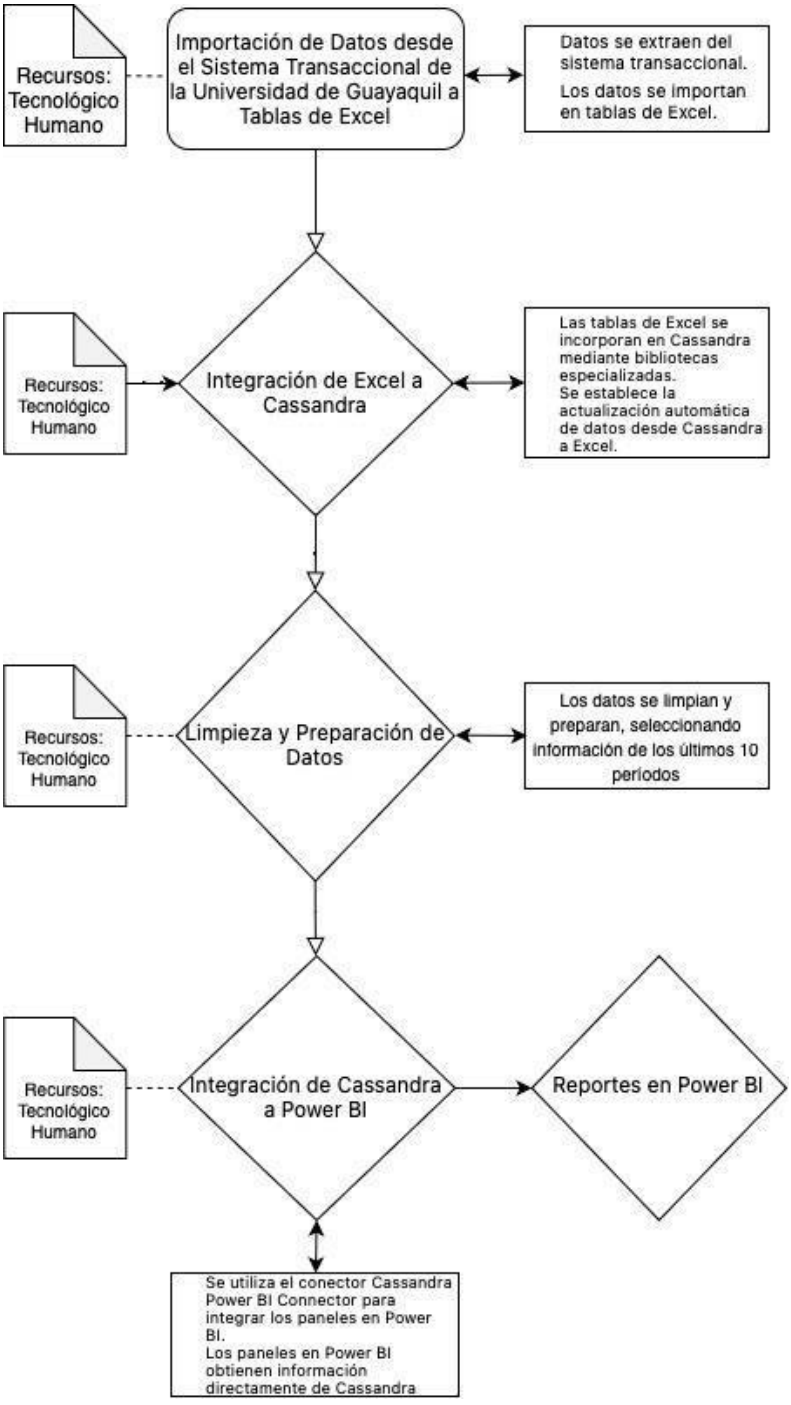
- **Análisis de Paralelos**
  - Comparación de los resultados entre diferentes paralelos de la misma materia.
  - Identificación de factores que pueden influir en el rendimiento de los estudiantes en diferentes grupos.

- **Análisis de Aprobación**
  - Evaluación del estado de aprobación en relación con el número de veces que se ha tomado la materia.
  - Identificación de patrones de mejora o deterioro en función de las repeticiones.

- **Tendencias Temporales**
  - Identificación de patrones estacionales o cíclicos en el rendimiento académico.
  - Visualización de tendencias a lo largo de diferentes años académicos.

Figura 11:

Pseudocódigo



2.6.6. NORMAS ISO

En la Universidad de Guayaquil, abordar la seguridad y privacidad de los datos se erige como una prioridad insoslayable. Se buscará salvaguardar celosamente la confidencialidad, integridad y accesibilidad de los datos bajo gestión. La solidez de este proyecto reposará en la solícita consideración de la seguridad y privacidad de los datos, así como la EGSI, lineamientos de la, así como la Ley Orgánica de Protección de Datos Personales. La implementación juiciosa de salvaguardias de seguridad desempeñará un papel crucial al asegurar tanto la confiabilidad como la integridad ética en la administración de los registros estudiantiles. Así, el éxito global del presente proyecto hallará cimiento en esta premisa irrevocable de protección y consideración de los datos sensibles, así como una implementación adecuada de medidas de seguridad contribuirá a garantizar la confiabilidad y la ética en el manejo de los datos de los estudiantes (Holguín Quimis, 2020).

Considerando que se manejarán datos delicados, incluyendo el desempeño académico de los alumnos, se adoptará una medida de seguridad clave. Se implementará la generación de identificadores únicos (proceso de Anonimización) para excluir información personal identificable no relevante para los datos personales. Esto asegurará que se resguarde la integridad y privacidad de la información al eliminar la visibilidad de elementos innecesarios. Esta práctica cobra vital importancia durante la extracción y carga de datos desde el sistema transaccional a Apache Cassandra. En este proceso, se seguirán escrupulosamente las pautas establecidas en un Estándar de Gestión de Seguridad de la Información (EGSI).

El EGSI es un marco integral que aborda la seguridad de la información desde una perspectiva holística. Proporciona una guía detallada para la gestión de riesgos, la protección de activos y la garantía de la confidencialidad, integridad y disponibilidad de los datos. Este enfoque reconoce la importancia de la colaboración entre las partes interesadas y promueve la conciencia de seguridad en toda la organización. En el entorno del presente proyecto, donde los datos son el activo más valioso, el EGSI se vuelve esencial para mitigar riesgos como el acceso no autorizado y la pérdida de datos. De igual manera se aplicaría lineamientos de normas ISO 27000, la norma ISO 27001 establece los requisitos para un sistema de gestión de seguridad de la información (SGSI) (Tonysé de la Rosa, 2021).

Al adoptar estos estándares, podríamos identificar, evaluar y tratar los riesgos de seguridad de manera eficiente. La norma no solo ayuda a establecer procesos sólidos de seguridad de la información, sino que también demuestra el compromiso de la organización con la protección de los datos en un entorno.

Figura 12:

Normas ISO 27001.

Nota: Obtenido de la web. Elaborado por Alex Campbell.

Se establecerá roles y permisos de acceso adecuados en las bases de datos. Solo el personal permitido deberá tener acceso a los datos, y se debe implementar un sistema de autenticación sólido para garantizar que solo las personas autorizadas puedan interactuar con los datos. Además, es crucial asegurarse de que el manejo de datos cumpla con las regulaciones y leyes de privacidad de datos aplicables, como la ley orgánica de protección de datos (LOPD) en Ecuador.

En un mundo donde los datos personales se recopilan y utilizan en una escala sin precedentes, la privacidad se ha convertido en una preocupación primordial. Las leyes de protección de datos, como la Ley Orgánica de Protección de Datos Personales, buscan salvaguardar los derechos y la privacidad de los individuos en relación con el procesamiento de sus datos personales.

Estas leyes establecen pautas para la recopilación, almacenamiento y uso de datos personales, así como la obligación de obtener el consentimiento informado de los titulares de los datos, en el presente proyecto se establecerá lineamientos estrictos para dicho cumplimiento.

Con los datos seleccionados lo exportaremos a un data-frame, el cual va a ser utilizado para poder realizar el análisis respectivo en Python. Con los datos obtenidos producto del análisis mediante machine learning se obtendrá una tabla de resultados que será importada en Cassandra, la misma que, se integrará nuevamente con Power BI para la generación del dashboard de resultados. Es importante recordar que esta integración va a requerir habilidades de programación y conocimiento técnico para establecerla adecuadamente. Además, debemos considerar al transferirlos entre plataformas.

2.6.7. ARQUITECTURA DEL MODELO

Para el proyecto presentado usaremos el análisis predictivo de datos, donde se recopilará y procesarán los datos de por lo menos diez periodos anteriores para encontrar patrones que sean útiles en la predicción de la cantidad de estudiantes por materia que aprobarán el curso actual. Se usará técnicas estadísticas de modelización, BIG DATA y machine learning para extraer datos históricos y realizar pronósticos.

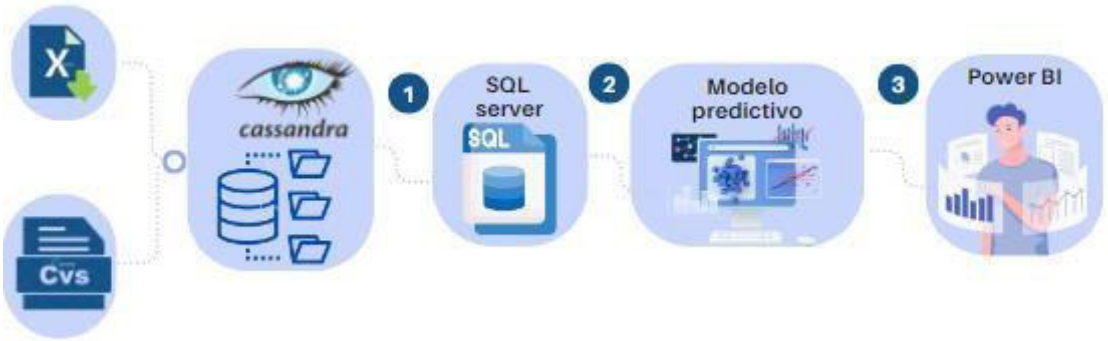
Los datos recopilados serán sometidos a un proceso de limpieza con el fin de evitar duplicados indeseados, o datos con información faltante, y también se analizará periodos atípicos como los producidos durante la pandemia del 2020, y algún otro tipo de información que no aporte significativamente a la resolución del problema. Este modelo puede ayudar a prever la cantidad de cursos realmente necesarios en cada semestre y materia, evitando de esta manera problemas en la planificación académica, optimizando los recursos disponibles.

Para nuestro proceso se usará una infraestructura ON-PREMISE que será implementada inicialmente en una máquina virtual y luego se buscará colocarla en un alojamiento de los servidores de la universidad con el fin de tener el mayor control sobre los datos, principalmente aquellos datos que puedan afectar a las leyes del procesamiento de datos personales.

Entre las herramientas a usar disponemos de los ficheros en formato *.xls y *.csv, que serán tratadas con el gestor de bases de datos Cassandra, luego se usara lenguajes de manipulación de datos y modelización con herramientas de Big-Data como Python, para la visualización de resultados se realizara paneles de Dashboard en Power BI.

Figura 13:

Arquitectura Cassandra



2.6.7.1. MEJORES PRÁCTICAS DE MANEJO DE DATOS

En la Universidad de Guayaquil, abordar la seguridad y privacidad de los datos se erige como una prioridad insoslayable. Inicialmente, se identificará y catalogará todos los datos personales que se utilizarán en el proyecto. Esto incluirá cualquier información que pueda identificar directa o indirectamente al estudiante, como nombres, direcciones de correo electrónico, números de identificación, etc. Se buscará salvaguardar celosamente la confidencialidad, integridad y accesibilidad de los datos bajo gestión. La solidez de este proyecto reposará en la solícita consideración de la seguridad y privacidad de los datos, así como la EGSI, lineamientos de la, así como la Ley Orgánica de Protección de Datos Personales. La implementación juiciosa de salvaguardias de seguridad desempeñará un papel crucial al asegurar tanto la confiabilidad como la integridad ética en la administración de los registros estudiantiles.

Así, el éxito global del presente proyecto hallará cimiento en esta premisa irrevocable de protección y consideración de los datos sensibles, así como una implementación adecuada de medidas de seguridad contribuirá a garantizar la confiabilidad y la ética en el manejo de los datos de los estudiantes. Considerando que se manejarán datos delicados, incluyendo el desempeño académico de los alumnos, se adoptará una medida de seguridad clave. Se implementará la generación de identificadores únicos (proceso de Anonimización) para excluir información personal identificable no relevante para los datos personales. Esto asegurará que se resguarde la integridad y privacidad de la información al eliminar la visibilidad de elementos innecesarios. Esta práctica cobra vital importancia durante la extracción y carga de datos desde el sistema transaccional a Apache Cassandra. En este proceso, se seguirán escrupulosamente las pautas establecidas en un Estándar de Gestión de Seguridad de la Información (EGSI).

El EGSI es un marco integral que aborda la seguridad de la información desde una perspectiva holística. Proporciona una guía detallada para la gestión de riesgos, la protección de activos y la garantía de la confidencialidad, integridad y disponibilidad de los datos. Este enfoque reconoce la importancia de la colaboración entre las partes interesadas y promueve la conciencia de seguridad en toda la organización. En el entorno del presente proyecto, donde los datos son el activo más valioso, el EGSI se vuelve esencial para mitigar riesgos como el acceso no autorizado y la pérdida de datos. De igual manera se aplicaría lineamientos de normas ISO 27000, la norma ISO 27001 establece los requisitos para un sistema de gestión de seguridad de la información (SGSI). Al adoptar estos estándares, podríamos identificar, evaluar y tratar los riesgos de seguridad de manera eficiente. La norma no solo ayuda a establecer procesos sólidos de seguridad de la información, sino que también demuestra el compromiso de la organización con la protección de los datos en un entorno (Haris, 2018).

Además, se obtendrá el consentimiento informado por parte de la Dirección de carrera de Telemática ya que es información fundamental que se utilizarán en el proyecto. Se informará de manera clara y transparente sobre cómo se utilizarán sus datos, los propósitos de la investigación y cualquier otro detalle relevante.

2.6.7.2. ESCALABILIDAD

En el ámbito de la escalabilidad, se plantea la utilización de servicios en la nube que posibiliten una expansión horizontal. Esto implica que la infraestructura pueda ampliarse sin dificultad para gestionar volúmenes mayores de datos y tráfico. Asimismo, se recomienda la adopción de aplicaciones y servicios en contenedores, lo que simplifica la escalabilidad automática al dividir las cargas de trabajo en unidades más manejables. Para garantizar un funcionamiento eficaz, se aconseja emplear herramientas de monitorización en tiempo real que puedan identificar cuellos de botella y problemas de rendimiento, permitiendo la configuración de alertas para abordar los desafíos de manera proactiva.

2.6.8. KPIS

Los KPI’s o también conocidos como indicadores clave de desempeño, evaluarán el rendimiento de las actividades y procesos en relación con los objetivos estratégicos establecidos previamente en el presente proyecto. Nos enfocaremos en las características primordiales de los KPI’s, los cuales serán enfocados a alcanzables, alineados, medibles, relevantes, en un plazo establecido de tiempo y exactos o sencillos. Los KPIs supervisarán y medirán el rendimiento de los objetivos y de ser el caso, durante el proceso de diseño del presente modelo predictivo de aprobación de materias mediante la analítica de datos para la planificación y proyección de cursos por periodo académico (carrera de Telemática Universidad de Guayaquil), se irán evaluando para el ajuste o replanteo de los mismo.

Porcentaje de precisión de resultados (modelo supervisado de clasificación)

- **Definición:** El porcentaje de predicciones correctas que el modelo ha hecho en comparación con el número total de predicciones realizadas.
- **Frecuencia de Medición:** se mide después de cada ciclo de entrenamiento del modelo o cuando se evalúa el rendimiento del modelo en un conjunto de datos de prueba.
- **Meta:** Alcanzar un porcentaje de precisión de resultados del 90% o superior en un plazo determinado.
- **Impacto en el negocio:** Evaluando si el modelo contribuye a la reducción de la huella ecológica al optimizar la asignación de recursos.

Tiempo de entrega de resultados del modelo

- **Definición:** El tiempo promedio, en horas o minutos, que transcurre desde que se solicita una predicción o resultado al modelo hasta que se entrega al usuario o sistema que lo requiere.
- **Frecuencia de Medición:** se mide para cada solicitud de predicción o resultado, y se puede calcular el promedio diario, semanal o mensual.
- **Meta:** Reducir el tiempo de entrega de resultados del modelo en un 20% en los próximos seis meses.

Error de la predicción (predicción – número de cursos necesarios para un semestre)

- **Definición:** El valor numérico que representa la diferencia entre el número de cursos predichos por el modelo y el número real de cursos necesarios para un semestre.
- **Frecuencia de Medición:** va a medirse después de cada ciclo de predicción o semestralmente, dependiendo de la frecuencia con la que se realicen las predicciones.
- **Meta:** Reducir el error de la predicción en un 10% en los próximos seis meses.

Tiempo de procesamiento del modelo

- **Definición:** El tiempo promedio, en segundos o milisegundos, que un modelo necesita para procesar una solicitud o tarea.
- **Frecuencia de Medición:** se procederá a medirse para cada solicitud de procesamiento o tarea, y se puede calcular el promedio diario, semanal o mensual.
- **Meta:** Reducir el tiempo de procesamiento del modelo en un 10% en los próximos seis meses.

Porcentaje de efectividad del modelo proyectado

- **Definición:** El porcentaje de coincidencia entre los valores proyectados por el modelo y los valores reales o esperados.
- **Frecuencia de Medición:** Se mide según la periodicidad de las proyecciones o modelos semestral.
- **Meta:** Alcanzar un porcentaje de efectividad del modelo proyectado del 90% o superior en un plazo determinado.
- **Impacto en el negocio:** Optimización de recursos en la disminución de cursos cerrados y/o bajo número de estudiantes matriculados.

Ahorro proyectado post implementación

- **Definición:** Ahorro proyectado referente a recursos implementados con la aplicación del modelo vs el gasto incurrido en semestres anteriores.
- **Frecuencia de Medición:** Debido a la naturaleza de los datos la frecuencia para este indicador es semestral, ya que ese es el periodo que dura el ciclo lectivo.
- **Meta:** Máximo un paralelo cerrado por semestre
- **Impacto en el Negocio:** Optimización de recursos.

2.6.9. METRICA DE VALOR

- **GASTO INCURRIDO POR CIERRE DE PARALELOS SIN EL MODELO PREDICTIVO 2023 CII**

Tabla 1:

Gasto de paralelos sin el modelo predictivo 2023 CII

RECURSOS	*VALOR HORA*	*HORAS*	*VALOR POR SEMESTRE*	*PARALELOS*	*VALOR X RECURSO*
Asesoría planificación	$8,75	7h	$61,75		$61.75
Recursos académicos (personal docente)	$8,75	96h	$840	7	$5.880
Recursos operativos (personal de servicios)	$4,17	48h	$200	7	$1.400
Recursos físicos (aula)	$12,5	96h	$1.200	7	$8.400
VALOR TOTAL					**$15.741,75**

- **GASTO PROYECTADO IMPLEMENTANDO EL MODELO PREDICTIVO 2024 CI**

Tabla 2:

Índices de paralelos con el modelo predictivo 2024 CI

RECURSOS	*VALOR HORA*	*HORAS*	*VALOR POR SEMESTRE*	*PARALELO*	*VALOR X RECURSO*
Asesoría planificación	$8,75	2h	$17,50		$17,50
Recursos académicos (personal docente)	$8,75	96h	$840	1	$840
Recursos operativos (personal de servicios)	$4,17	48h	$200	1	$200
Recursos físicos (aula)	$12,5	96h	$1.200	1	$1.200
VALOR TOTAL					**$2.257,50**

- **AHORRO PROYECTADO CON LA APLICACIÓN DEL MODELO PREDICTIVO**

Tabla 3:

Ahorro proyectado con la aplicación del modelo predictivo.

Gastos incurridos sin el modelo 2023 CII	$15.741,75
Gastos incurridos con el modelo 2024 CI	$2.257,50
Valor de ahorro proyectado	$13.484,25

2.6.10. ANÁLISIS PESTEL

Figura 14:

Análisis Pestel,

Nota: Obtenido de la web. Elaborado por grupo trevenque.

- **POLÍTICAS**

"LEY ORGANICA DE EDUCACION SUPERIOR, LOES, Registro Oficial Suplemento 298 de 12-oct.-2010

Art. 2.- Objeto. - Esta Ley tiene como objeto definir sus principios, garantizar el derecho a la educación superior de calidad que propenda a la excelencia interculturalidad, al acceso universal.

Art. 25.- Rendición anual de cuentas de fondos públicos. - Las instituciones del Sistema de Educación Superior deberán rendir cuentas de los fondos públicos recibidos en relación con sus fines, mediante el mecanismo que establezca la Contraloría General del Estado, en coordinación con el órgano rector de la política pública de educación superior, y conforme las disposiciones de la Ley que regula el acceso a la información." (República del Ecuador, 2018).

Política educativa y gubernamentales en el sector educativo pueden tener un impacto significativo en la universidad. Cambios en la legislación, financiamiento y regulaciones pueden afectar las operaciones y la estrategia de la universidad. Adicionalmente la estabilidad política en Ecuador es importante para el funcionamiento sin problemas de la universidad. Cualquier inestabilidad política podría afectar negativamente la seguridad y la inversión en educación superior.

- Cambio en las políticas de estado, referente al libre acceso a la Universidad.

- **ECONÓMICAS**

"REGLAMENTO DE DISTRIBUCION RECURSOS INSTITUCIONES EDUCACION SUPERIOR, Resolución del Consejo de Educación Superior 39

Art. 10.- Fórmula de distribución de recursos. - Para la distribución anual de las rentas o asignaciones del Estado a las universidades y escuelas politécnicas se empleará la siguiente fórmula:

Para las universidades y escuelas politécnicas del literal a) del artículo 3 de este Reglamento, el monto restante del rubro correspondiente a la gratuidad se distribuirá con los indicadores calculados con información exclusivamente de tercer nivel.

Art. 169.- h) Aprobar la fórmula de distribución anual de las rentas o asignaciones del Estado a las instituciones de educación superior y de los incrementos si es que los hubiere, las que constarán en el Presupuesto General del Estado, de acuerdo a los lineamientos de la presente Ley." (República del Ecuador, 2019)

Referente al ciclo económico la universidad puede verse afectada por los ciclos económicos de Ecuador. Durante períodos de recesión, el financiamiento y las matrículas pueden disminuir, mientras que, en épocas de crecimiento económico, pueden aumentar. Así mismo, el capital económico y fondos disponibles están estrechamente ligados a la partida presupuestaria provista por el gobierno del Ecuador.

- Mayor o menor asignación de recursos, provistos en relación con el número de estudiantes ingresados.

- **SOCIALES**

"REGLAMENTO GENERAL A LA LEY ORGÁNICA DE EDUCACIÓN INTERCULTURAL

Art. 222.- Evaluación del comportamiento. La evaluación del comportamiento de los estudiantes en las instituciones educativas cumple un objetivo formativo motivacional y está a cargo del docente de aula o del docente tutor. Se debe realizar en forma literal y descriptiva, a partir de indicadores referidos a valores éticos y de convivencia social, tales como los siguientes: respeto y consideración hacia todos los miembros de la comunidad educativa, valoración de la diversidad, cumplimiento con las normas de convivencia, cuidado del patrimonio institucional, respeto a la propiedad ajena, puntualidad y asistencia, limpieza, entre otros aspectos que deben constar en el Código de Convivencia del establecimiento educativo.

Art. 155.- Acceso al servicio educativo público. - Para el ingreso a las instituciones educativas públicas, la Autoridad Educativa Nacional establecerá el procedimiento de inscripción, asignación de cupos y matrícula, cumpliendo con el principio de acercar el servicio educativo a los usuarios." (República del Ecuador, 2015)

La demografía estudiantil y los cambios en la población estudiantil, como la cantidad de estudiantes en edad universitaria y sus preferencias educativas, pueden influir en la demanda de programas académicos, así también, los cambios en los valores culturales y sociales pueden afectar las áreas de investigación, los programas académicos y las políticas de diversidad de la universidad.

- Reducción de insatisfacción estudiantil, gracias a una mejor proyección y planificación de apertura de cursos.

- **TECNOLÓGICAS**

"POLÍTICA PARA LA TRANSFORMACIÓN DIGITAL DEL ECUADOR 2022-2025, Ministerio de Telecomunicaciones.

5.1. Infraestructura. 5.1.2. Penetración de Internet, porcentaje de hogares con acceso a internet 53.20%.

Porcentaje de personas que tienen teléfono inteligente, por grupos de edad de 16 a 24 años 74.90%." (República del Ecuador, 2022)

La adopción de nuevas tecnologías en la educación, como la educación en línea o la inteligencia artificial, puede cambiar la forma en que la universidad ofrece sus servicios. La infraestructura de TI de la universidad es crucial para el funcionamiento eficiente de sus operaciones académicas y administrativas. Durante la pandemia se adaptaron políticas educativas on-line para poder continuar con la educación durante los tiempos de confinamiento.

- Mejora en la predicción de proyección académica, como resultado de un correcto análisis de datos académicos históricos.
- Identificación temprana de riesgo de no aprobación (académico) acorde a datos, tipo de materia y tipo de docente.

- **AMBIENTALES**

"LEY DE GESTION AMBIENTAL, CODIFICACION, Registro Oficial Suplemento 418 de 10-sep-2004

e) Regular y promover la conservación del medio ambiente y el uso sustentable de los recursos naturales en armonía con el interés social." (República del Ecuador, 2004)
La universidad puede verse afectada por la creciente preocupación por la sostenibilidad y las regulaciones ambientales. Esto puede influir en las prácticas de gestión y en la oferta de programas referente al espacio físico disponible.

- Uso sostenible de recursos, debido a la optimización del número de cursos apertura dos, reduciendo consumo indebido de energía y espacios físicos.

- **LEGALES**

“LEY ORGÁNICA DE PROTECCIÓN DE DATOS PERSONALES, Quinto Suplemento del Registro Oficial No.459, 26 de Mayo 2021.

Art. 2.- Ámbito de aplicación material.• La presente ley se aplicará al tratamiento de datos personales contenidos en cualquier tipo de soporte, automatizados o no, así como a toda modalidad de uso posterior.

Ar. 37.- Seguridad de datos personales, Entre otras medidas, se podrán incluir las siguientes;

1.- Medidas de Anonimización, seudonomización o cifrado de datos personales.” (República del Ecuador, 2021)

Las leyes y regulaciones relacionadas con la educación superior pueden influir en la estructura y operación de la universidad. Las cuestiones legales relacionadas con la propiedad intelectual y los derechos de autor son importantes para la investigación y la propiedad de contenido académico.

- Protección de datos personales sobre la data obtenida para el análisis de la proyección de cursos a apertura.
- Calificación y acreditación de la carrera por medio de índices de cumplimiento estudiantil de aprobación y culminación de carrera.

2.6.11. PLANTEAMIENTO AGILE

El desarrollo de un modelo predictivo de aprobación de materias mediante la analítica de datos para la planificación y proyección de cursos por periodo académico en la carrera de Telemática es un proyecto que se beneficiaría significativamente de la aplicación de una metodología Agile. La metodología Agile permitirá una mayor flexibilidad, adaptabilidad y colaboración a lo largo del proyecto, lo que es esencial cuando se trabaja en un entorno de desarrollo de un modelo (código) para la analítica de datos (Hadida, 2020). Se optará por la implementación de la metodología Scrum, uno de los enfoques de Agile más populares y efectivos para la gestión de proyectos de este tipo. A continuación, se plantea una metodología Agile utilizando Scrum como base para el presente proyecto.

1. **Definición del Equipo:**

El proyecto reúne un equipo multidisciplinario que incluya a analistas de datos, expertos en la carrera de Telemática de la Universidad de Guayaquil, desarrolladores, un Scrum Máster y un Product Owner. Cabe recalcar que el equipo estará formado por dos personas que actualmente laboran en la Universidad de Guayaquil, las que aportarán un gran conocimiento para un correcto desarrollo.

- **Product Owner:** Expertos en la carrera de Telemática o un representante de la facultad que tenga un profundo conocimiento de los requisitos y objetivos del proyecto.
- **Scrum Máster:** Un facilitador que ayudará al equipo a seguir los principios de Scrum.
- **Equipo de Desarrollo**: Grupo interdisciplinario que incluya analistas de datos, ingenieros de software, y otros profesionales necesarios para el proyecto.

Figura 15:

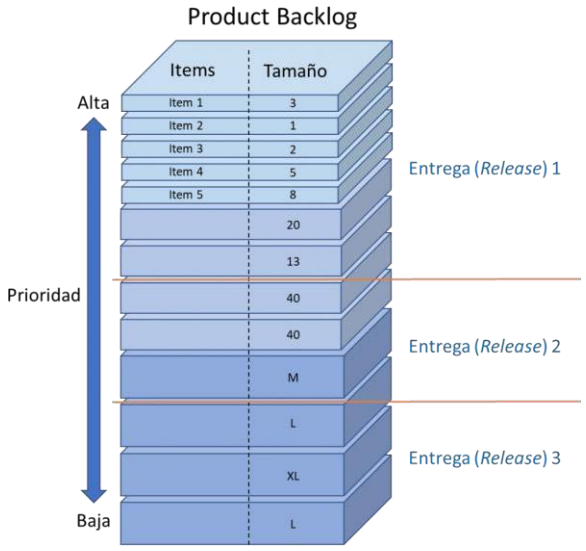
Equipo Scrum.

2. **Creación del Backlog del Producto**

El Product Owner trabajará con el equipo para identificar y priorizar las características clave del modelo predictivo y la planificación de cursos. Estas características se registrarán en el Backlog del Producto, que se mantendrá actualizado a medida que se obtenga más información y se refinen los requisitos.

Figura 16:

Backlog del Producto.



Nota: Obtenido de la web. Elaborado por muy ágil.

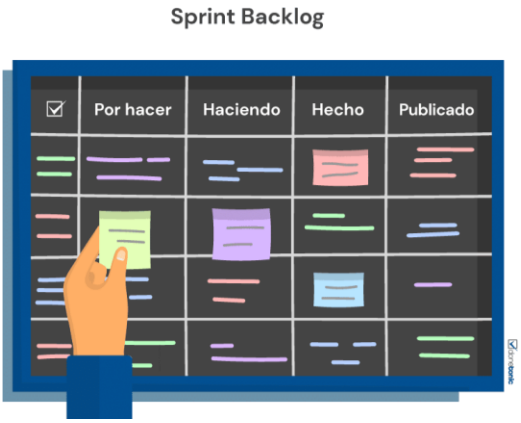
3. Planificación del Sprint

El equipo realizará reuniones de planificación de Sprint para seleccionar un conjunto de elementos del Backlog del Producto que se abordarán en un Sprint específico. Cada Sprint debe tener una duración fija de dos semanas.

- **Sprint 0 - Preparación:** En este primer Sprint, el equipo se enfocará en establecer la infraestructura, seleccionar las herramientas y definir las normas y procedimientos que se utilizarán a lo largo del proyecto.
- **Planificación del Primer Sprint:** Se seleccionará un conjunto inicial de elementos del Product Backlog que se abordarán en el primer Sprint. Así mismo se mantendrán reuniones de planificación del Sprint donde el equipo planificará cómo completar las tareas seleccionadas durante el Sprint, definiendo el alcance y los criterios de aceptación.

Figura 17:

Product Backlog y Sprint.



Nota: Obtenido de la web. Elaborado por canva.

4. Desarrollo y Análisis

Durante el Sprint, el equipo trabajará en el desarrollo del modelo predictivo y las funcionalidades relacionadas aplicadas a la Universidad de Guayaquil. Se realizarán depuración de la base de datos para su posterior análisis y se ajustarán los modelos según sea necesario para mejorar la precisión de proyección de materias. Se llevarán a cabo reuniones diarias de seguimiento (Daily Standup) para mantener al equipo informado sobre el progreso y abordar cualquier impedimento.

- **Reuniones diarias**: realiza reuniones diarias cortas para mantener al equipo informado sobre el progreso, identificar obstáculos y adaptar el plan si es necesario.

Figura 18:

Daily Standup,

Nota: Obtenido de la web. Elaborado por canva.

5. Iteración

El ciclo se repetirá a lo largo del proyecto, refinando y ampliando el modelo predictivo y las funcionalidades en cada Sprint. La retroalimentación constante y la adaptabilidad son clave para el éxito del proyecto, lo cual permite ajustar los recursos disponibles.

6. Entrega Incremental:

A medida que se completen Sprints, se entregará valor incremental a los potenciales usuarios del modelo (Universidad de Guayaquil), lo que permite una implementación gradual de las capacidades del sistema. Esta metodología Agile basada en Scrum permitirá una gestión efectiva del proyecto del modelo predictivo de aprobación de materias mediante la analítica de datos para la planificación y proyección de cursos por periodo académico en la carrera de Telemática, una respuesta ágil a los cambios en los requisitos y una colaboración continua entre el equipo y las partes, lo que debería conducir al desarrollo exitoso del modelo predictivo y la planificación de cursos para la carrera de Telemática.

Figura 19:

Metodología Agile basada en SCRUM.

Nota: Obtenido de la web. Elaborado por nimble.

2.6.12. PLANIFICACIÓN DE RECURSOS

En base a la planificación de recursos, nos abarcará 1623 horas de las cuales, el ingeniero de datos cargará los datos desde la base de la Universidad hasta el almacén intermedio de datos (Cassandra) y también procesará los datos para la fase de modelamiento, lo cual tomará 348 horas, el analista de datos en las 660 horas revisará y transformará la información obtenida para que quede lista para el modelamiento y realizará los informes y los cuadros de mando KPIs, el

científico de datos en las 280 horas en encargará del modelamiento de datos para obtener las predicciones esperadas, el gobierno de datos en las 185 horas es el encargado durante todo el proceso de verificar que se cumpla lo estipulado con la Ley de Protección de Datos en el Ecuador y el ingeniero de DevOps con las 150 horas juntará toda la información en una aplicación amigable para una mejor comprensión de los resultados obtenidos. Los recursos están distribuidos entre diseño técnico, implementación e implantación.

Tabla 4:

Cronograma

	Estimación ***Total***	***Ingeniero de datos*** ***Horas***	***Analista de datos*** ***Horas***	***Científico de datos*** ***Horas***	***Gobierno de datos*** ***Horas***	***Ingeniero de DevOps*** ***Horas***
Diseño Técnico	133					
Estrategia de Ingreso de datos						
** Cargar datos desde CSV*	5	5				
Modelado de Información						
** Almacén intermedio (Landin Zone)*	9	9				
** Información Específica (Data Marts)*	4	4				
Gobierno de datos						
** Análisis de Leyes (Seguridad)*	5				5	
** Calidad*	10				10	
Explotación						
** informes*	50		50			
** KPIs*	50		50			
Implementación	970					
Ingreso de datos						
** Cargar datos desde CSV*	150	150				
Transformación de Información						
** Procesamiento*	400	100	100	200		
Explotación						
** informes*	100		100			
** Cuadros de mando KPIs*	200		200			
Gobierno de datos						
** Seguridad*	50				50	
** Calidad*	50				50	
** Metadatos*	20				20	
Implantación	520					
Pruebas	170	50	50	50	20	
Documentación	120	30	30	30	30	
Implantación	150					150
Formación	80		80			
TOTAL	1623	348	660	280	185	150

CAPITULO III

3.1. ANÁLISIS DE RESULTADO

El proyecto es un esfuerzo para mejorar la planificación de los cursos a dictarse por periodo académico teniendo una proyección analizada mediante técnicas de Machine Learning para predecir cuantos estudiantes aprueban los cursos en que se matriculan.

3.1.1. DATOS UTILIZADOS

Se toma los datos desde la plataforma institucional de la Universidad de Guayaquil la que viene dada por periodo y en distintas tablas de Excel, estos datos son almacenados en cassandra y una vez almacenado todas las tablas se procede a depurar los datos.

En la depuración se establece cuales columnas tienen los datos que realmente vamos a utilizar y realizando un estudio de los datos disponibles se concluyó que los campos que podrían tener mejor rendimiento serían.

- Periodo,
- Identificación
- Nivel_mat
- Cod_materia
- Ccdocen
- [asistencia 1p]
- [promedio p1]
- [promedios asistencia]
- [promedio final]
- Sexo
- Edad
- Discapacidad

Con esta información se eliminó aquellas filas que no contenían la información requerida, y partiendo de lo establecido en el reglamente general de calificaciones de la universidad se calculó cuales estudiantes tienen el curso aprobado si para aprobar se debe tener un mínimo de 70% de [PROMEDIOS ASISTENCIA] y un [PROMEDIO FINAL] mayor o igual a 7.00, se realizó mediante consultas CQLSH y se guardó en una columna con el nombre de “ESTADOAP”.

Con esta información se calculó otras columnas que concluimos que eran necesarias para mejorar la precisión de la predicción que son la Eficiencia de aprobación de materias por parte del estudiante la cual es la relación entre las materias aprobadas y las materias tomadas por cada estudiante en forma porcentual, así mismo se evaluó la eficiencia de aprobación por

docente que se calcula tomando la cantidad de estudiantes que aprueban la materia por cada docente en relación al nuero de estudiantes que toma su materia, y de la misma forma se procedió a calcular la eficiencia de aprobación por materia en la que se consideró por cada materia la relación entre estudiantes que aprueban dicha materia versus el número de estudiantes que toma dicha materia, estos datos se colocan en las columnas “EfiDocen, EfiEstud. EfiMater”.

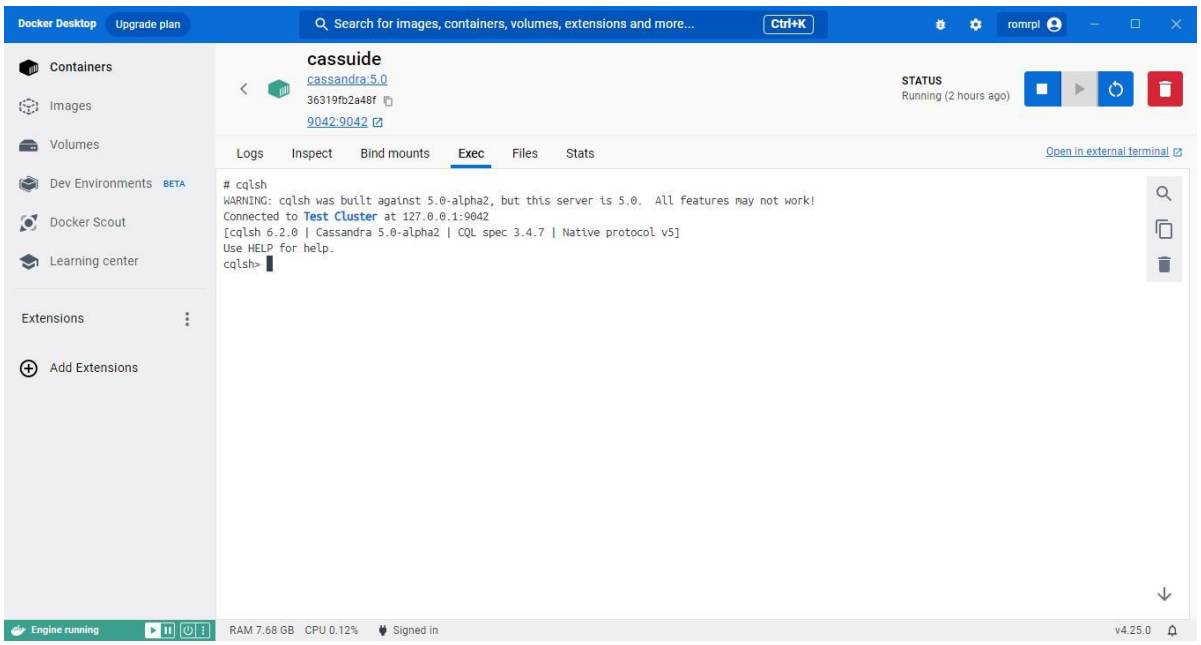
3.1.2. ALMACENAMIENTO DE DATOS

El uso de Cassandra para almacenar los datos es una elección razonable, ya que es una base de datos NoSQL escalable que puede manejar grandes volúmenes de datos y es adecuada para aplicaciones con altas tasas de escritura y lectura, como sistemas de seguimiento académico.

- Inicializa el contenedor de Docker con la imagen de Cassandra versión 5.0

Figura 20:

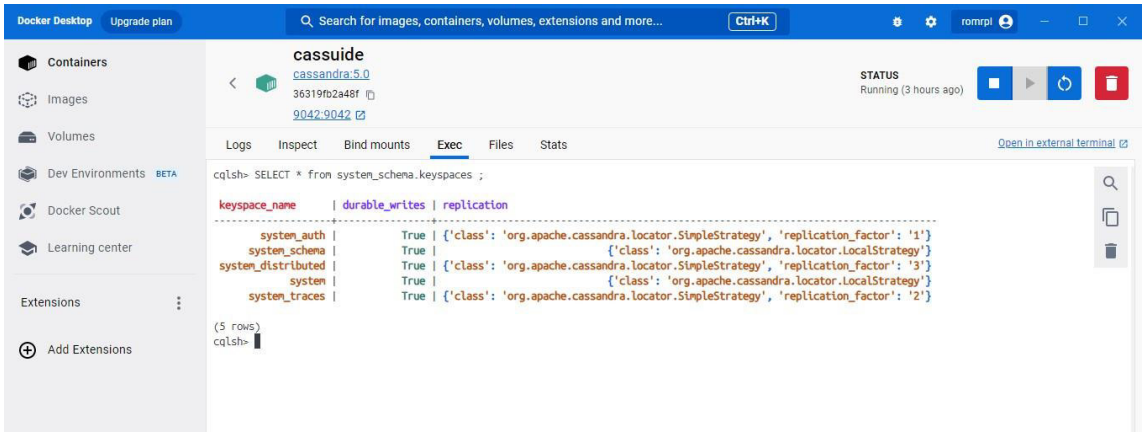
Docker con Cassandra



- Se verifica que no existan Keyspaces antes de iniciar el proceso

Figura 21:

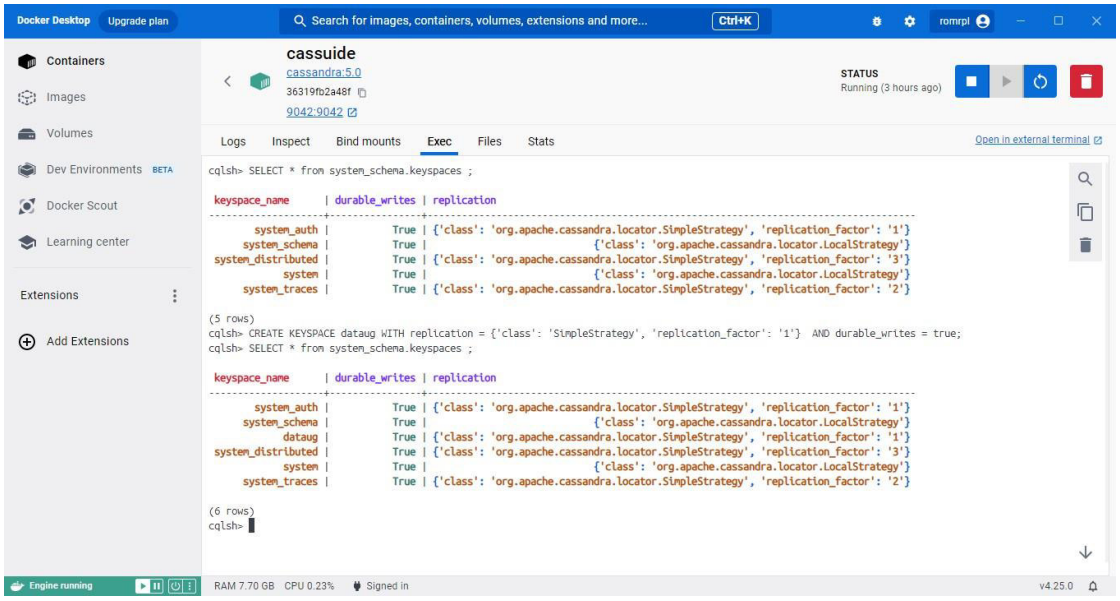
Existencia de Keyspace



- Creación de esquema de base llamado “dataug”

Figura 22:

Creación de la base de datos “DATAUG”



3.1.3. MODELOS UTILIZADOS

Los modelos empleados para realizar este proyecto fueron, "Modelo de Árbol de Decisión", "Random Forest" y "Regresión Logística", todos son ampliamente utilizados en problemas de clasificación supervisada, y son adecuados para este tipo de proyecto. La precisión obtenida en todos fue del 91% lo que da un resultado bastante sólido y sugiere que el proyecto puede tener una buena tasa de éxito en predecir quiénes aprobarán las materias. En vista de los resultados se escogió usar únicamente el "Modelo de Árbol de Decisión".

3.1.4. PYTHON

Python es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML). Los desarrolladores utilizan Python porque es eficiente y fácil de aprender, además de que se puede ejecutar en muchas plataformas diferentes. El software Python se puede descargar gratis, se integra bien a todos los tipos de sistemas y aumenta la velocidad del desarrollo. (AWS, 2023).

Figura 23:

Logo Python.

Nota: Obtenido de la web. Elaborado por Python, Inc.

.

3.1.5. LIBRERÍAS

3.1.5.1. PANDA

La biblioteca de software de código abierto Pandas está diseñada específicamente para la manipulación y el análisis de datos en el lenguaje Python. Es potente, flexible y fácil de usar. Gracias a Pandas, por fin se puede utilizar el lenguaje Python para cargar, alinear, manipular o incluso fusionar datos. El rendimiento es realmente impresionante cuando el código fuente del back-end está escrito en C o Python.

El nombre Pandas es en realidad una contracción del término Panel Data para series de datos que incluyen observaciones a lo largo de varios periodos de tiempo. La biblioteca se creó como herramienta de alto nivel para el análisis en Python. (DataScientest, pandas, 2023).

Figura 24:

Logo de pandas.

Nota: Obtenido de la web. Elaborado por Pandas, Inc.

3.1.5.2. NUMPY

NumPy es un proyecto de código abierto que permite la computación numérica con Python. Fue creado en 2005 basándose en los primeros trabajos de las bibliotecas Numeric y Numarray, siempre será un software 100% de código abierto y gratuito para todos. Se publica bajo los términos liberales de la licencia BSD modificada. Se desarrolla abiertamente en GitHub, a través del consenso de NumPy y la comunidad científica más amplia de Python. Para obtener más información sobre nuestro enfoque de gobernanza, consulte nuestro Documento de gobernanza. (NUMPY, 2023)

Figura 25:

Numpy.

Nota: Obtenido de la web. Elaborado por Numpy, Inc.

3.1.5.3. MATPLOTLIB

Matplotlib es una biblioteca completa para crear visualizaciones estáticas, animadas e interactivas en Python, hace que las cosas fáciles sean fáciles y las difíciles posibles. (Hunter, 2023).

Figura 26:

Matplotlib.

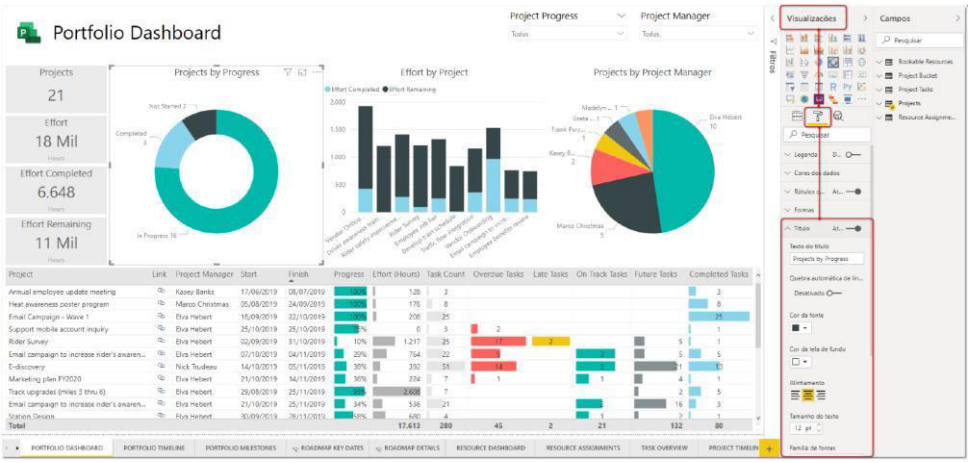
Nota: Obtenido de la web. Elaborado por matplotlib. Inc.

3.1.5.4. POWER BI

Power BI es una plataforma unificada y escalable de inteligencia empresarial (BI) con funciones de autoservicio apta para grandes empresas. Conéctese a los datos, visualícelos e incorpore sin problemas objetos visuales en las aplicaciones que usa todos los días. (Microsoft, 2023).

Figura 27:

Power BI.



Nota: Obtenido de la web. Elaborado por Microsoft Support.

3.1.5.5. SKLEARN

Es una biblioteca de Python que proporciona acceso a versiones eficaces de muchos algoritmos comunes. También proporciona una API propia y estandarizada. Por tanto, una de las grandes ventajas de Scikit-Learn es que una vez que se entiende el uso básico y su sintaxis para un tipo de modelo, cambiar a un nuevo modelo o algoritmo es muy sencillo. (DataScientest, 2023).

Figura 28:

Logo de Scikit Learm.

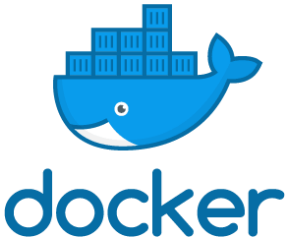
Nota: Obtenido de la web. Elaborado por David Cournapeau.

3.1.6. DOCKER

Docker es un proyecto de código abierto que automatiza el despliegue de aplicaciones dentro de contenedores de software, proporcionando una capa adicional de abstracción y automatización de virtualización de aplicaciones en múltiples sistemas operativos. Docker proporciona un conjunto de herramientas de desarrollo, servicios, contenido confiable y automatizaciones, utilizados individualmente o en conjunto, para acelerar la entrega de aplicaciones seguras. (DOCKER, 2023)

Figura 29:

Logo de Docker.



Nota: Obtenido de la web. Elaborado por Docker, Inc.

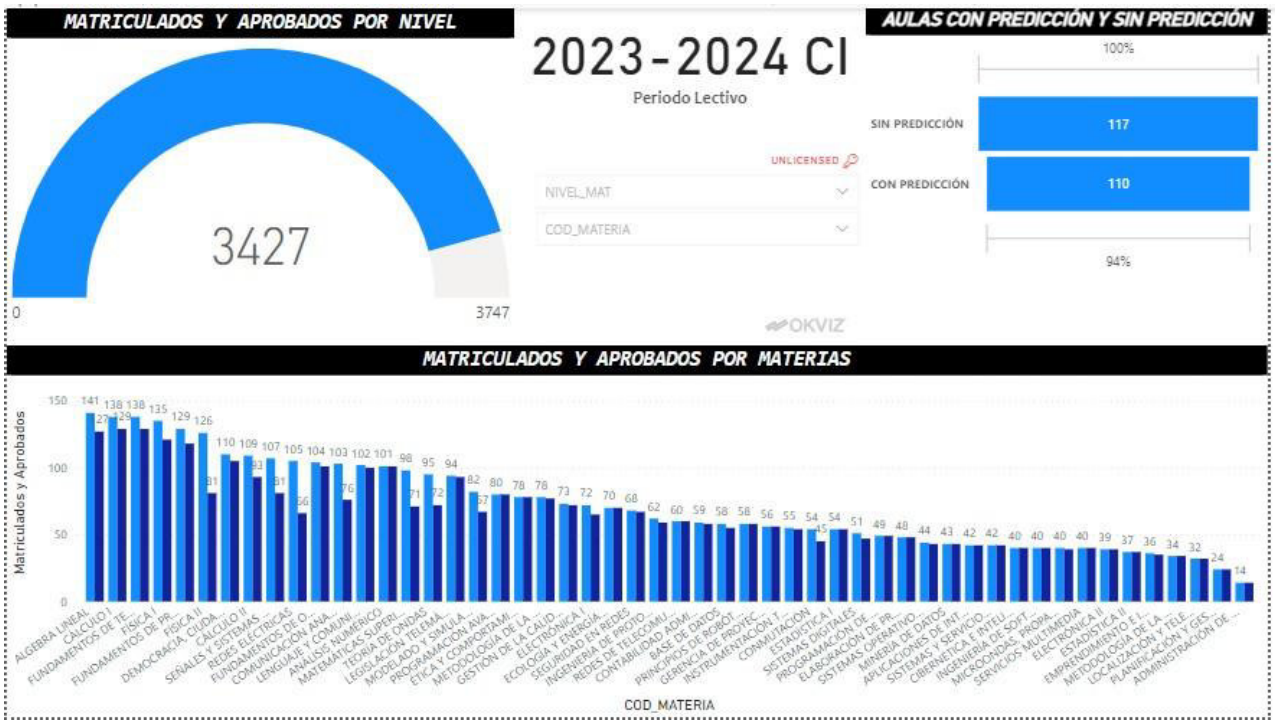
3.1.7. VISUALIZACIÓN DE RESULTADOS

Para la visualización de datos se escogió Power BI por ser una poderosa herramienta y de fácil uso, lo que la hace muy adecuada para presentar los resultados de manera efectiva. Con Power BI, es posible crear paneles y gráficos interactivos que permiten a los usuarios analizar y comprender los resultados de manera intuitiva.

En la gráfica siguiente, se ofrece una representación visual del conjunto total de estudiantes matriculados por nivel y materias en contraste con los estudiantes que lograron la aprobación durante el semestre 2023-2024 CI. En la parte inferior de la gráfica, se desglosa exhaustivamente el total de estudiantes matriculados y aprobados en relación con el conjunto de materias, presentando esta información de manera clara a través de barras individuales. Como parte del análisis integral, se proyecta el número global de aulas que se abrirán en el semestre, destacando tanto la predicción generada por el modelo como la situación proyectada sin la aplicación de la predicción. Este enfoque holístico permitirá una comprensión más profunda de las dinámicas académicas y de planificación, proporcionando información valiosa para la toma de decisiones estratégicas.

Figura 30:

Información general Power BI.

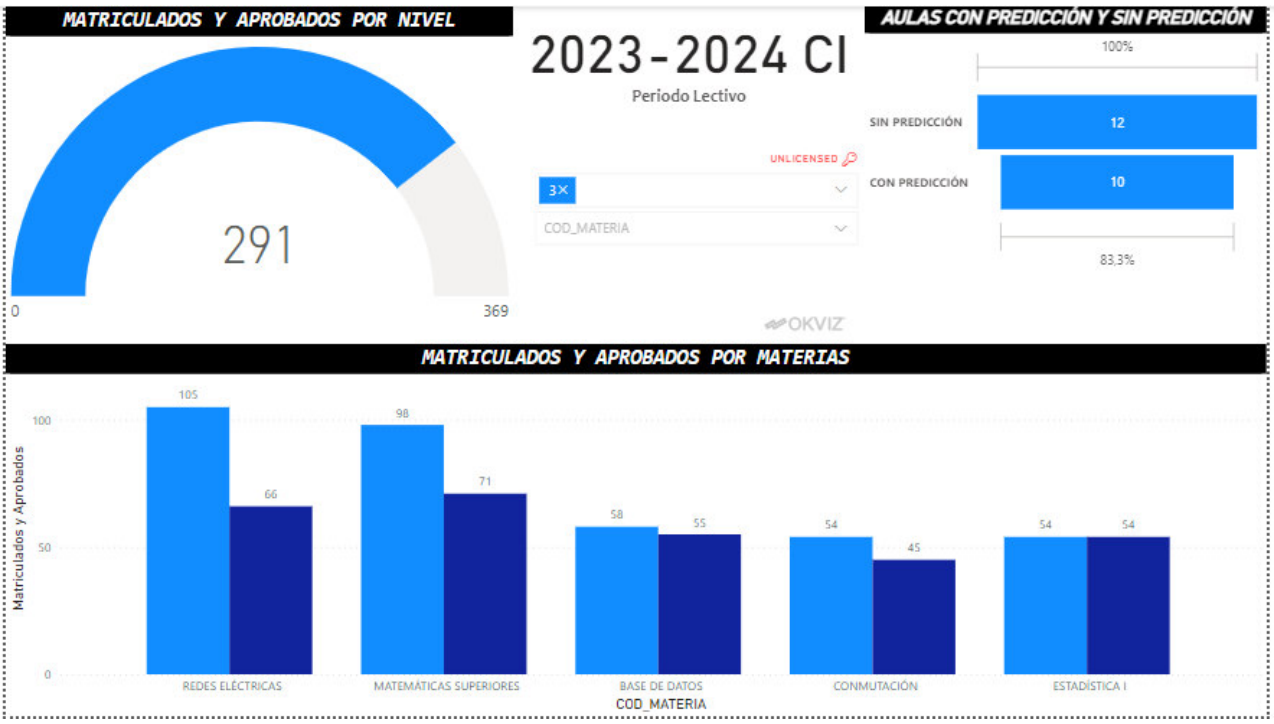


A continuación, se presentan los diagramas correspondientes al nivel 3, donde se observa que se han matriculado 369 estudiantes, de los cuales 291 han sido aprobados. Además, se proporciona un análisis detallado de las aulas en este nivel: sin la aplicación de la predicción, se abrieron 12 aulas, mientras que, con la implementación del modelo predictivo, se proyecta la apertura de 10 aulas.

En los gráficos de barras asociados a este nivel, se detallan las materias impartidas, destacando el número de estudiantes matriculados y aprobados en cada una. Es importante señalar que la elección de este nivel específico se basa en su complejidad, siendo uno de los semestres que presenta mayores desafíos en cuanto a la planificación académica y la apertura de aulas. Este enfoque permitirá un análisis más preciso de las dinámicas que influyen en la gestión académica, contribuyendo a una planificación más efectiva y ajustada a las demandas de este nivel particular.

Figura 31:

Power BI por nivel

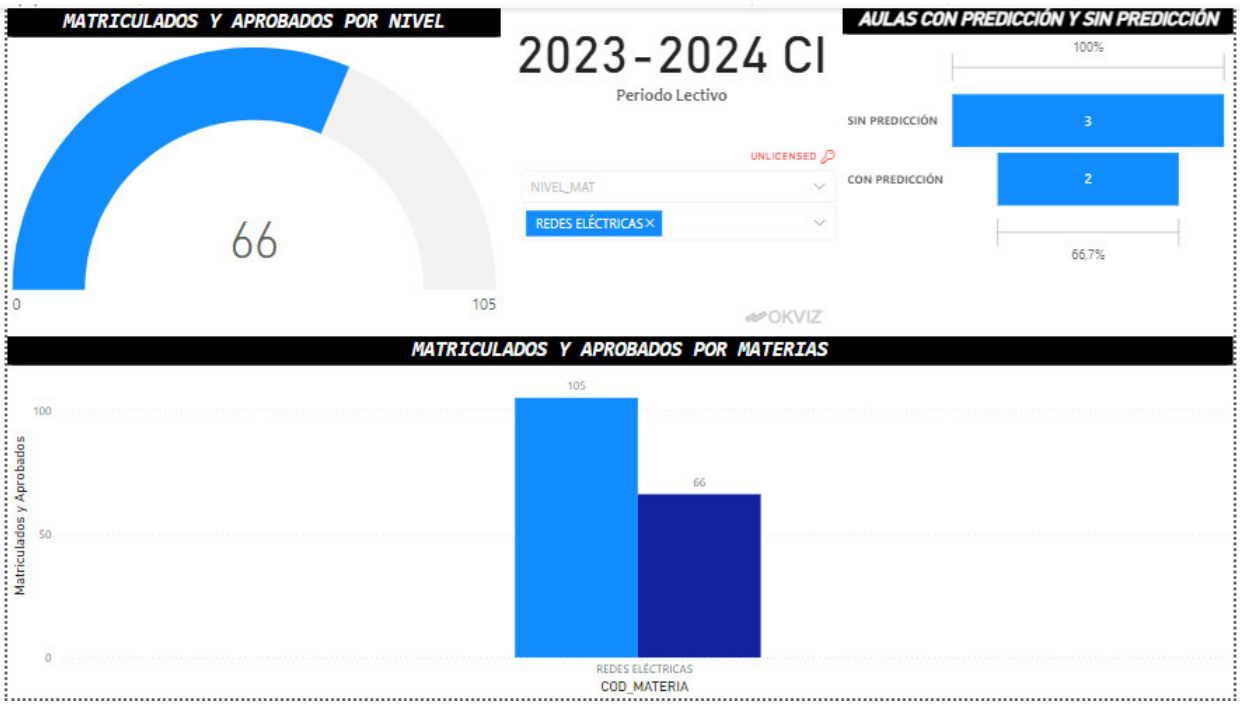


Concluyendo la revisión de los resultados, la imagen presenta una proyección específica para la materia de redes eléctricas. En este análisis, observamos que se matricularon 105 estudiantes, de los cuales 66 lograron aprobar, como se detalla en la gráfica correspondiente. La apertura de aulas asociadas a esta materia, sin la aplicación de la predicción, fue de 3 aulas, mientras que, con la implementación del modelo predictivo, la proyección indica la apertura de 2 aulas.

Es crucial destacar que la materia de redes eléctricas se considera especialmente compleja, y hemos elegido examinarla específicamente debido a la notable diferencia en los resultados entre la aplicación del modelo y la ausencia de este. Este enfoque nos permite destacar la eficacia del modelo predictivo en situaciones de complejidad académica, proporcionando una valiosa referencia para futuras planificaciones y toma de decisiones.

Figura 32:

Power BI por materia



3.1.8. EVALUACIÓN DE RESULTADOS

La precisión del 91% es un resultado positivo y sugiere que el modelo es capaz de identificar con éxito a la mayoría de los estudiantes que aprobarán las materias. Sin embargo, es importante realizar una evaluación más profunda de los resultados.

Se deben considerar métricas adicionales, como la sensibilidad (recall) y la especificidad, para comprender mejor el rendimiento del modelo, especialmente en la identificación de falsos positivos y falsos negativos.

3.1.9. OPTIMIZACIÓN

Para mejorar aún más el modelo se planifica buscar otra información que sea relevante como situación económica, tiempo de movilización, Institución de donde proviene el estudiante, pero aún no se dispone de esta información y se planteará obtenerla por medio de encuestas para futuras mejoras, se pueden explorar diferentes técnicas, como la selección de características, la ingeniería de características y la optimización de hiperparámetros. Además, es importante monitorear el rendimiento del modelo a medida que se recopilan más datos y ajustar el modelo en consecuencia.

3.1.10. USO POTENCIAL

Los resultados de este proyecto pueden ser utilizados por la universidad para identificar a los estudiantes que pueden necesitar apoyo adicional, ya sea a través de tutorías, programas de asesoramiento académico o intervenciones específicas para mejorar su rendimiento académico.

3.1.11. ÉTICA

Es importante considerar la ética en el uso de estos modelos. Los resultados no deben utilizarse para discriminar a los estudiantes, sino para proporcionarles apoyo y recursos adicionales.

CAPITULO IV

4.1. CONCLUSIONES

- Se identificaron variables predictoras significativas que afectan la aprobación de materias, como el rendimiento previo, eficiencia académica del estudiante, porcentaje de asistencia a clases, evaluaciones parciales, entre otros factores relevantes para el diseño del modelo.

- El diseño de este modelo puede ayudar a la carrera Telemática a planificar la oferta académica de manera más eficiente, mejorando la satisfacción estudiantil y evitando cierres innecesarios de cursos.

- Este modelo ayudará generando una mayor eficiencia operativa interna, lo cual representa un ahorro significativo en recursos, tanto de personal como de equipamiento.

- Se realizó tres tipos de modelos, árbol de decisión, Random Forest y regresión lineal, el cual se eligió el árbol de decisiones por ser el que tenía un alto porcentaje en la precisión de la predicción.

- Revisando el histórico de cursos abiertos en el período 2023 CII y comparando con el modelo se pudo comprobar la eficacia de utilizarlo ya que se proyectó 110 paralelos usando el modelo y 117 utilizando la planificación de forma tradicional.

4.2. RECOMENDACIONES

- La implementación de un modelo predictivo de aprobación de materias mediante analítica de datos puede ser una herramienta valiosa para mejorar la planificación de cursos en la carrera de Telemática de la Universidad de Guayaquil. Sin embargo, es fundamental abordar estos esfuerzos de manera ética y continua, centrándose en el beneficio de los estudiantes y la calidad de la educación.

- Se deben establecer políticas internas sólidas de privacidad y ética para garantizar que la recopilación y el uso de datos cumplan con las regulaciones vigentes.

- Es fundamental seguir recopilando y agregando datos que se consideren relevantes para mejorar la precisión del modelo con el tiempo, al ir semestre a semestre alimentando el modelo con data ya generada por el mismo y eliminando datos atípicos, se tendrá en un futuro mejores resultados.

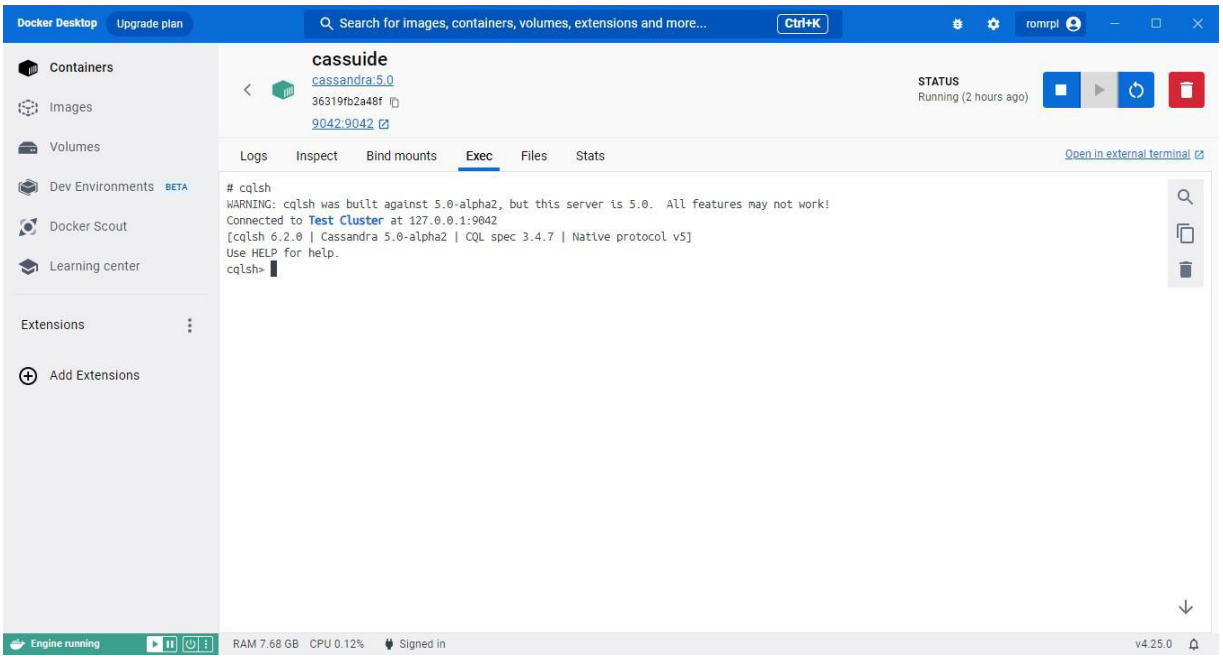
- Se recomienda realizar un plan piloto una vez optimizado el modelo en las carreras de la Facultad Ingeniería Industrial para una posterior implementación a nivel de Universidad, ya que podría proveernos un ahorro considerable en recursos.

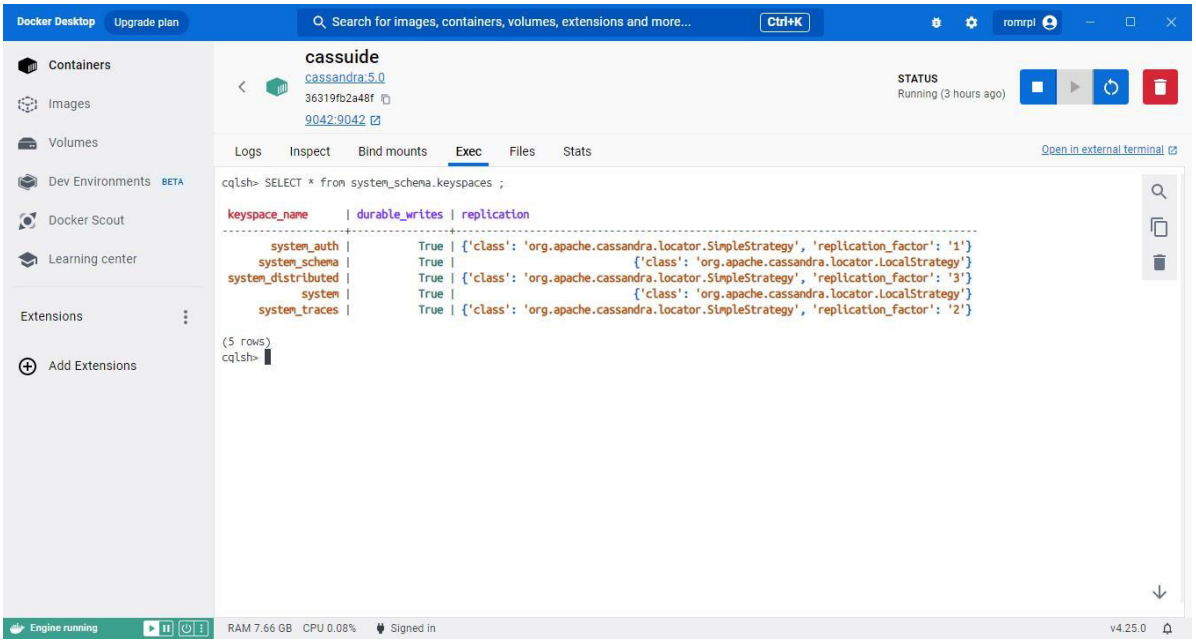
- Implementar la obtención de datos adicionales, en los procesos de matriculación para mejorar el modelo, como el nivel socioeconómico del estudiante, tiempo de traslado, estado civil, educación previa, puntaje de ingreso a la Universidad.

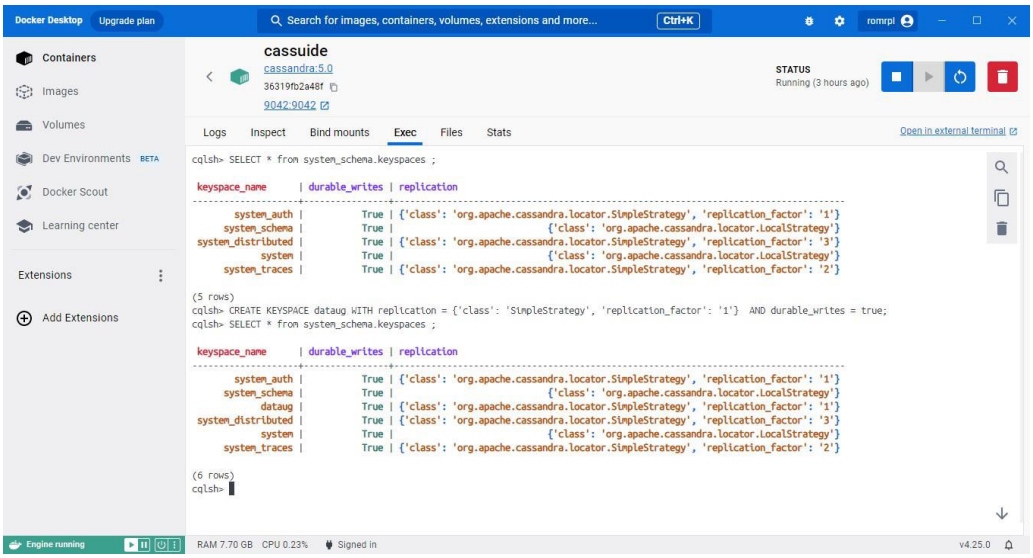
5. APÉNDICE

Link Git Hub: https://github.com/MixiJoselyne/ModeloPredictivo

DOCKER Y CASSANDRA







MODELOS PREDICTIVOS

MODELO: ÁRBOL DE DECISIÓN

1. Instalar las librerías requeridas.

- python -m pip install --upgrade pip
- python -m pip install Pillow
- python -m pip install -U matplotlib
- python -m pip install scipy
- python -m pip install pandas
- python -m pip install scikit-learn
- python -m pip install openpyxl

```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import tree
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score
```

```
def CondicionAprobado(fila): # Definición de la función que asigna la condición del estado del estudiante x materia
    if fila["ASISTENCIA"]>=70 and fila["PROM FINAL"]>=7.0:
        cap="Aprobado"
    else:
        cap="Reprobado"
    return cap
```

```
def convertird(dato,tipo):
    convertido=""
    if tipo == "text":
        convertido=str("'" + str(dato) + "'")
    elif tipo == "int":
        convertido=str(dato)
    elif tipo == "date":
        convertido=str("'" + str(datetime.strftime(datetime.strptime(str(dato), '%d/%m/%Y %H:%M:%S'),'%Y-%m-%d')) + "'")
    elif tipo == "bigint":
        convertido=str(dato)
    elif tipo == "float":
        convertido=str(dato)
    else:
        convertido=str("'" + str(dato) + "'")
    return convertido
```

```
ArchiDirec=os.getcwd()+"\\data\\calif\\"
# Obtener la lista de archivos en la carpeta
archivos = os.listdir(ArchiDirec)

# Conecta con el cluster de Cassandra
cluster = Cluster(['localhost'])
session = cluster.connect('dataug')
# Preparar la tabla en Cassandra
session.execute('''
    CREATE TABLE IF NOT EXISTS calificaciones (
    periodo TEXT, identificacion TEXT, nivel INT, cod_materia BIGINT, cod_tcalifica INT, docente TEXT, nota DOUBLE,
    PRIMARY KEY (periodo, identificacion, nivel, cod_materia, cod_tcalifica, docente))''')
```

```
for archivo in archivos:
    print(ArchiDirec + archivo)
    # Lee el archivo Excel
    df = pd.read_excel( ArchiDirec + archivo)
    # Insertar datos en Cassandra
    for index, row in df.iterrows():
        # Obtener los datos con el formato acorde a los campo en cassandra
        tipoCal=row['COD_TCALIFICA']
        if (tipoCal==151615  or tipoCal== 151635 or tipoCal== 151625 or tipoCal== 151633):
            addrow =convertird(row['PERIODO'],"text")  + "," + convertird(row['IDENTIFICACION'],"text") + "," + convertird(re.search(r'\d+',row['GRUPO'][9:11]).group(),"int") \
                + "," + convertird(row['COD_MATERIA'],"bigint") + "," + convertird(row['COD_TCALIFICA'],"int") + "," + convertird(row['DOCENTE'],"text")[:10] + "," + convertird(row['NOTA'],"flo
            #print(addrow)
            session.execute('INSERT INTO calificaciones (periodo,identificacion, nivel, cod_materia, cod_tcalifica,docente,nota) VALUES (' + addrow +')')
```

```
# Ruta de datos=carpeta actual\\data\\califTest : Aquí están los datos del ciclo actual para la predicci{on
ArchiDirec=os.getcwd()+"\\data\\califTest\\"
# Obtener la lista de archivos en la carpeta
archivos = os.listdir(ArchiDirec)

# Preparar la tabla en Cassandra
session.execute('''
    CREATE TABLE IF NOT EXISTS calificacionesT (
    periodo TEXT, identificacion TEXT, nivel INT, cod_materia BIGINT, cod_tcalifica INT, docente TEXT, nota DOUBLE,
    PRIMARY KEY (periodo, identificacion, nivel, cod_materia, cod_tcalifica, docente))''')
```

```
for archivo in archivos:
    print(ArchiDirec + archivo)
    # Lee el archivo Excel
    df = pd.read_excel( ArchiDirec + archivo)
    # Insertar datos en Cassandra
    for index, row in df.iterrows():
        # Obtener los datos con el formato acorde a los campoe en cassandra
        tipoCal=row['COD_TCALIFICA']
        if (tipoCal==151615  or tipoCal== 151635):
            addrow =convertird(row['PERIODO'],"text")  + "," + convertird(row['IDENTIFICACION'],"text") + "," + convertird(re.search(r'\d+',row['GRUPO'][9:11]).group(),"int") \
                + "," + convertird(row['COD_MATERIA'],"bigint") + "," + convertird(row['COD_TCALIFICA'],"int") + "," + convertird(row['DOCENTE'],"text")[:10] + "'," + convertird(row['NOTA'],"flo
            #print(addrow)
            session.execute('INSERT INTO calificacionesT (periodo,identificacion, nivel, cod_materia, cod_tcalifica,docente,nota) VALUES (' + addrow +')')
```

```
cluster = Cluster(['localhost'])
session = cluster.connect('dataug')

query = "SELECT * FROM calificaciones"
result = session.execute(query)

# Convertir los resultados de la consulta a un DataFrame de pandas
df = pd.DataFrame(result)
# Cerrar la conexión
cluster.shutdown()
# Hacer la transformación de datos
DataCalif = df.pivot_table(index=['periodo' , 'nivel', 'cod_materia', 'docente' ,'identificacion'],
                           columns='cod_tcalifica',
                           values='nota',
                           aggfunc='first').reset_index()

# Cambiar nombres de columnas por valores mas amigables
DataCalif.rename(columns={'periodo': 'PERIODO'}, inplace=True)
DataCalif.rename(columns={'nivel': 'NIVEL'}, inplace=True)
DataCalif.rename(columns={'cod_materia': 'COD_MATERIA'}, inplace=True)
DataCalif.rename(columns={'docente': 'DOCENTE'}, inplace=True)
DataCalif.rename(columns={'identificacion': 'IDENTIFICACION'}, inplace=True)
DataCalif.rename(columns={151615: 'ASISTENCIA 1P'}, inplace=True)
DataCalif.rename(columns={151635: 'PROMEDIO P1'}, inplace=True)
DataCalif.rename(columns={151625: 'PROMEDIOS ASISTENCIA'}, inplace=True)
DataCalif.rename(columns={151633: 'PROMEDIO FINAL'}, inplace=True)
```

```
DataCalif['EfiDocen']  = None
DataCalif['EfiEstud']  = None
DataCalif['EfiMater']  = None

# Aplicar la condición de aprobación de materia
DataCalif['ESTADOAP']  = DataCalif.apply(CondicionAprobado, axis = 1)

# Eliminar columnas no necesarias
DataCalif = DataCalif.drop('PROMEDIOS ASISTENCIA', axis=1)
DataCalif = DataCalif.drop('PROMEDIO FINAL', axis=1)
```

```python
# Encontrar el ultimo periodo ingresado
periodos = DataCalif.groupby('PERIODO')['ESTADOAP'].count()
periodos.sort_index(inplace=True)
uperiodo = periodos.index[-1]
#print(uperiodo)

# Calcular eficiencia de aprobación de estudiantes
EfiMatri = DataCalif.groupby('IDENTIFICACION')['ESTADOAP'].count()
EfiAprob = DataCalif.groupby('IDENTIFICACION')['ESTADOAP'].sum()
Efiestud = EfiAprob / EfiMatri
DataCalif['EfiEstud']=(DataCalif.apply(lambda x: Efiestud[x['IDENTIFICACION']], axis=1))
# Calcular eficiencia de aprobación de materia
EfiMatri2 = DataCalif.groupby('COD_MATERIA')['ESTADOAP'].count()
EfiAprob2 = DataCalif.groupby('COD_MATERIA')['ESTADOAP'].sum()
EfiMater = EfiAprob2 / EfiMatri2
DataCalif['EfiMater']=(DataCalif.apply(lambda x: EfiMater[x['COD_MATERIA']], axis=1))
# Calcular eficiencia de aprobación de docentes
EfiMatri3 = DataCalif.groupby('DOCENTE')['ESTADOAP'].count()
EfiAprob3 = DataCalif.groupby('DOCENTE')['ESTADOAP'].sum()
EfiDocen = EfiAprob3 / EfiMatri3
DataCalif['EfiDocen']=(DataCalif.apply(lambda x: EfiDocen[x['DOCENTE']], axis=1))
# exportar los datos listos para analisis y modelado
ArchivoPX = os.getcwd()+"\\data\\datafin.xlsx"
DataCalif.to_excel(ArchivoPX, index=False)
print("Estudiante", Efiestud.head())
print("Materias", EfiMater.head())
print("Docentes",EfiDocen.head())
```

```
Estudiante IDENTIFICACION
1004941322    0.000000
107111510     1.000000
107348385     1.000000
1104419831    1.000000
1104643489    0.978723
Name: ESTADOAP, dtype: float64
Materias COD_MATERIA
1516001    0.738255
1516002    0.772643
1516003    0.518333
1516004    0.891258
1516005    0.918159
Name: ESTADOAP, dtype: float64
Docentes DOCENTE
060337539    0.707941
090688189    0.840491
090792195    0.921569
090833926    0.926829
090956599    0.936270
Name: ESTADOAP, dtype: float64
```

```
# Promedio de eficiencia para estudiante
DataCalifUP = DataCalif[(DataCalif['PERIODO'] == uperiodo) & (DataCalif['NIVEL'] == 1)]
EfiMatriU = DataCalifUP.groupby(['IDENTIFICACION'])['ESTADOAP'].count()
EfiAprobU = DataCalifUP.groupby(['IDENTIFICACION'])['ESTADOAP'].sum()
EfiestudU = EfiAprobU / EfiMatriU
EfiPPEstN1 = EfiestudU.mean()
EfiMatriU.drop(EfiMatriU.index, inplace=True)
EfiAprobU.drop(EfiAprobU.index, inplace=True)
EfiestudU.drop(EfiestudU.index, inplace=True)
DataCalifUP.drop(DataCalifUP.index, inplace=True)

# Promedio de eficiencia para materias
DataCalifUP = DataCalif[(DataCalif['PERIODO'] == uperiodo)]
EfiMatriU = DataCalifUP.groupby(['COD_MATERIA'])['ESTADOAP'].count()
EfiAprobU = DataCalifUP.groupby(['COD_MATERIA'])['ESTADOAP'].sum()
EfiMatU = EfiAprobU / EfiMatriU
EfiPPMat = EfiMatU.mean()
EfiMatriU.drop(EfiMatriU.index, inplace=True)
EfiAprobU.drop(EfiAprobU.index, inplace=True)
DataCalifUP.drop(DataCalifUP.index, inplace=True)

# Promedio de eficiencia para docentes
DataCalifUP = DataCalif[(DataCalif['PERIODO'] == uperiodo)]
EfiMatriU = DataCalifUP.groupby(['DOCENTE'])['ESTADOAP'].count()
EfiAprobU = DataCalifUP.groupby(['DOCENTE'])['ESTADOAP'].sum()
EfiDocU = EfiAprobU / EfiMatriU
EfiPPDoc = EfiDocU.mean()
EfiMatriU.drop(EfiMatriU.index, inplace=True)
EfiAprobU.drop(EfiAprobU.index, inplace=True)
DataCalifUP.drop(DataCalifUP.index, inplace=True)
```

```
cluster = Cluster(['localhost'])
session = cluster.connect('dataug')

query = "SELECT * FROM calificacionest"
result = session.execute(query)

# Convertir los resultados de la consulta a un DataFrame de pandas
df = pd.DataFrame(result)

# Cerrar la conexión
cluster.shutdown()

# Hacer la transformación de datos
DataCalifT = df.pivot_table(index=['periodo' , 'nivel', 'cod_materia', 'docente' ,'identificacion'],
                            columns='cod_tcalifica',
                            values='nota',
                            aggfunc='first').reset_index()
```

```python
# Cambiar nombres de columnas por valores mas amigables
DataCalifT.rename(columns={'periodo': 'PERIODO'}, inplace=True)
DataCalifT.rename(columns={'nivel': 'NIVEL'}, inplace=True)
DataCalifT.rename(columns={'cod_materia': 'COD_MATERIA'}, inplace=True)
DataCalifT.rename(columns={'docente': 'DOCENTE'}, inplace=True)
DataCalifT.rename(columns={'identificacion': 'IDENTIFICACION'}, inplace=True)
DataCalifT.rename(columns={151615: 'ASISTENCIA 1P'}, inplace=True)
DataCalifT.rename(columns={151635: 'PROMEDIO P1'}, inplace=True)

# Calcular eficiencia de aprobación de estudiante

#df_actualizado = pd.merge(df, serie_identificacion, left_on='columna1', right_index=True, how='left')
EfiDocen.name = "EfiDocen"
DataCalifT=pd.merge(DataCalifT,EfiDocen, left_on='DOCENTE', right_index=True, how='left')
Efiestud.name = "EfiEstud"
DataCalifT=pd.merge(DataCalifT,Efiestud, left_on='IDENTIFICACION', right_index=True, how='left')
EfiMater.name = "EfiMater"
DataCalifT=pd.merge(DataCalifT,EfiMater, left_on='COD_MATERIA', right_index=True, how='left')
DataCalifT['ESTADOAP']  = None

DataCalifT = DataCalifT.fillna({'EfiDocen': EfiPPDoc, 'EfiEstud': EfiPPEstN1, 'EfiMater': EfiPPMat})
ArchivoPX=os.getcwd()+"\\data\\dataTest.xlsx"
DataCalifT.to_excel(ArchivoPX, index=False)
#print(type(Efiestud))
```

```python
# Cargar datos que se usaran como entrenamiento, al final se cargaran los datos para las predicciones

ArchiDirec=os.getcwd()+"\\data\\" # Ruta de datos=carpeta actual\\data\\
Archivo=ArchiDirec + "datatrain.xlsx" # Archivo con la información en formato xlsx
DataCalif=pd.read_excel(Archivo)

# El DataFrame usado es lo suficientemente grande y contiene suficientes opciones de aprobado o reprobado
# Localizamos los duplicados segun los campos que deben dar datos únicos, escogiendo la primera aparición de ser el caso
DataDuplicada = DataCalif.duplicated(subset = ["PERIODO","IDENTIFICACION","NIVEL_MAT","COD_MATERIA","CDOCEN"], keep = 'first')

# Buscamos y eliminamos las filas que tengan valores nulos en el campo "ASISTENCIA" y "PROM FINAL", y "PROM P1" ya que se necesita la información completa en estos campos
DataCalif.drop(DataCalif[(DataCalif["ASISTENCIA 1P"].isnull() )].index, inplace=True)
DataCalif.drop(DataCalif[(DataCalif["PROMEDIO P1"].isnull() )].index, inplace=True)
```

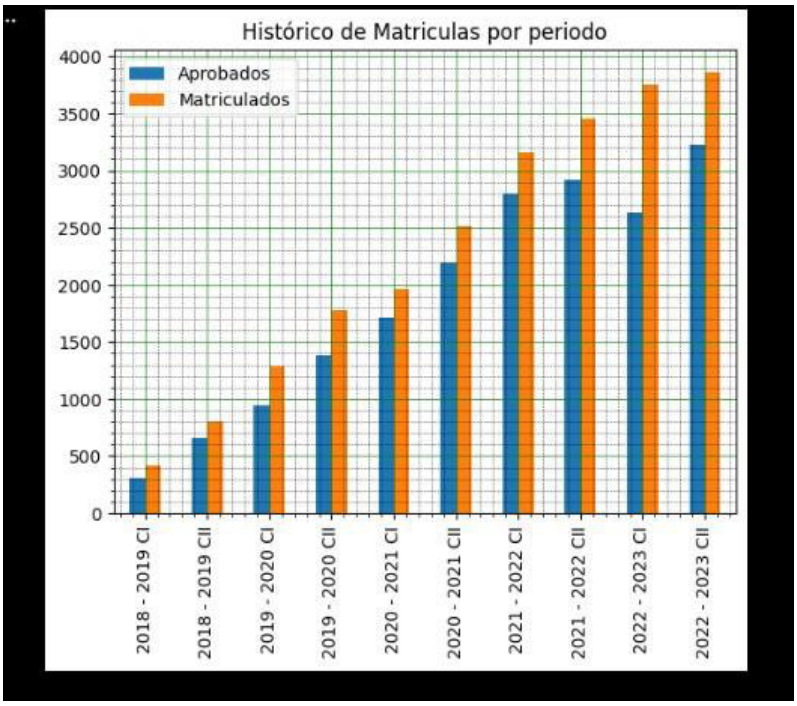
[2] Python

```python
# Se genera la condición de estado con "Aprobado" si un estudiante cumple con tener un promedio mayor o igual a 7.0 y Tener asistencia mayor o igual a 70%
# para esto se usa una función definida con el nombre de CondicionAprobado que revisa las columnas necesarias y define si el estado es Aprobado o no

def CondicionAprobado(fila): # Definición de la función que asigna la condición del estado del estudiante x materia
    if fila["ASISTENCIA"]>=70 and fila["PROM FINAL"]>=7.0:
        cap="Aprobado"
    else:
        cap="Reprobado"
    return cap

#DataCalif['ESTADOAP']  = DataCalif.apply(CondicionAprobado, axis = 1)  # Se añade una columna denominada "ESTADOAP" donde se almacena el estado de aprobación del estudiante x materia
```

[3] Python

```python
DataPeriodos=np.array(DataCalif["PERIODO"].unique()) # se busca las filas unicas en el campo "PERIODO"
c2=[] # Almacena la cantidad de estudiantes que tienen la condición de Aprobado
c3=[] # Almacena la cantidad total de estudiantes que hay en el PERIODO
DataPeriodos.sort() # ordena los periodos

for DataPeriodo in DataPeriodos: # Se preparan los datos que se mostrarán en la grafica
    c2.append(DataCalif.apply(lambda x: x['ESTADOAP'] == 1 and x["PERIODO"] == DataPeriodo, axis=1).sum() )
    c3.append(DataCalif.apply(lambda x: x["PERIODO"] == DataPeriodo, axis=1).sum())

DataGrafi=pd.DataFrame({'Aprobados':c2, 'Matriculados':c3}, index=DataPeriodos)
DataGrafi.plot(kind='bar', title='Histórico de Matriculas por periodo')
plt.minorticks_on() # Para activar las grillas en el gráfico
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
#plt.show()
```

[4] Python

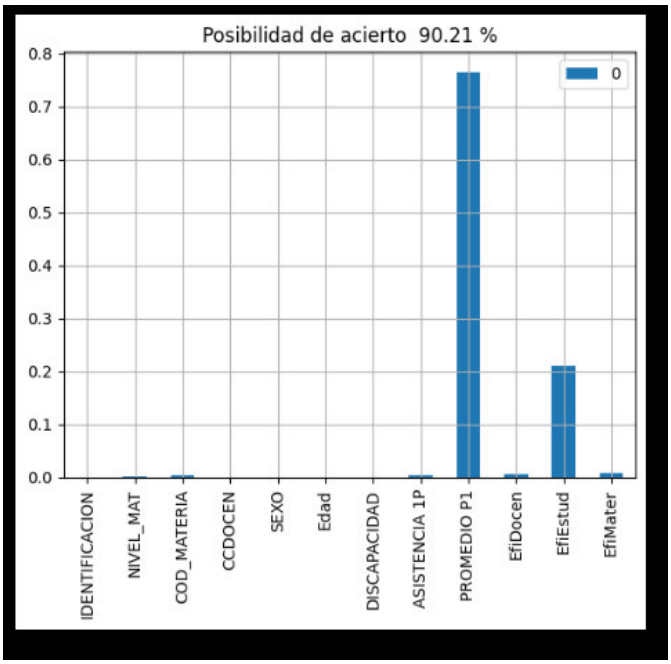



```
# Se escoje las columnas que se van a utilizar para el analisis predictivo
# (variables independientes) y se asignan a la variable "x"
x= DataCalif.copy().iloc[:,1:13]
x.info()
# Se escoje la columna de resultados (variable dependiente)
# se asignan a la variable "y"
y= DataCalif["ESTADOAP"]

# Se divide la muestra en datos de entrenamiento y de prueba#, random_state=42)
# el atributo  random_state= es para que siempre se escoja los mismos datos
X_train, X_test, y_train, y_test = train_test_split(x, y, train_size = 0.75)

# Verificamos que los datos de nuestro DataFrame son numericos
# x = x.astype(float)
# x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 22989 entries, 0 to 23001
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   IDENTIFICACION  22989 non-null  int64
 1   NIVEL_MAT       22989 non-null  int64
 2   COD_MATERIA     22989 non-null  int64
 3   CCDOCEN         22989 non-null  int64
 4   SEXO            22989 non-null  int64
 5   Edad            22989 non-null  int64
 6   DISCAPACIDAD    22989 non-null  int64
 7   ASISTENCIA 1P   22989 non-null  float64
 8   PROMEDIO P1     22989 non-null  float64
 9   EfiDocen        22989 non-null  float64
 10  EfiEstud        22989 non-null  float64
 11  EfiMater        22989 non-null  float64
dtypes: float64(5), int64(7)
memory usage: 2.3 MB
```

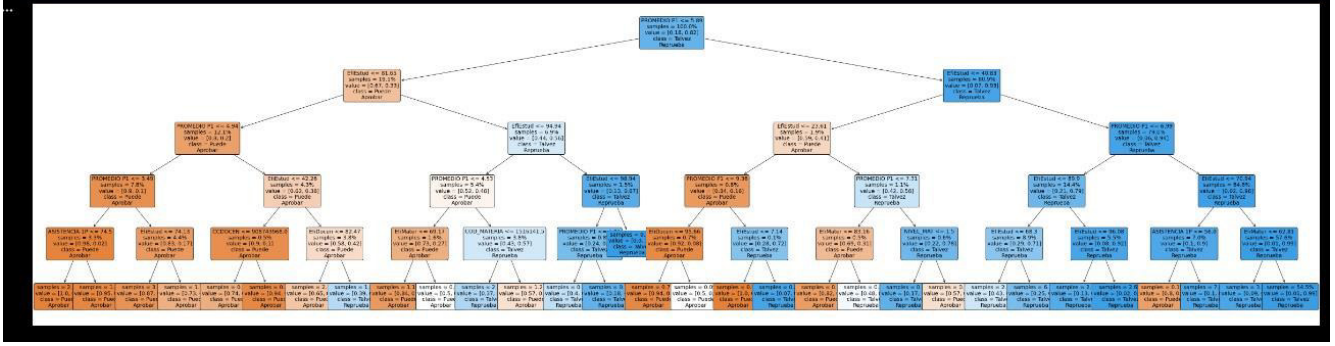
```
DataArbolDecis = tree.DecisionTreeClassifier(max_depth=5) # Generamos el arbol de desiciones con una profundidad de 5 niveles
DataArbolDecis = DataArbolDecis.fit(X_train, y_train)
yPredicha = DataArbolDecis.predict(X_test) # Ejecutamos la prediccion de prueba
DataPreciPredi=accuracy_score(y_test, yPredicha) # accuray_score nos permite conocer el valor porcentual de aciertos en la prediccion
DataMatriPredi=confusion_matrix(y_test, yPredicha,labels = (1, 0)) # La matriz de confusión nos enseña cuantas veces una clase fue asiganada correctamente
# precision_score muestra el porcentaje de posibilidad que la prediccion sea correcta
precision_score(y_test, yPredicha, pos_label = 1)
precision_score(y_test, yPredicha, pos_label = 0)
DataPrediPosib = pd.DataFrame(DataArbolDecis.feature_importances_, index = x.columns) # Obtenemos aquellas variables que son más importantes en nuestro arbol de decisión
DataPrediPosib.plot(kind="bar",title=f"Posibilidad de acierto  {round(DataPreciPredi*100,2)} %",grid=True) # Grafica de los indices que son relevantes para la prediccion
```

```
<Axes: title={'center': 'Posibilidad de acierto  90.21 %'}>
```




```
# Graficamos nuestro arbol de decision con una profundidad de 5 niveles

DataGrafoArbol = plt.figure(figsize=(60,15))
_ = tree.plot_tree(DataArbolDecis,max_depth=5,impurity = False, rounded = True, proportion=True,
                   feature_names = x.columns, precision=2,
                   class_names = {0:"Puede\nAprobar",1:"Talvez\nReprueba"},
                   filled = True,
                   fontsize = 14)
```

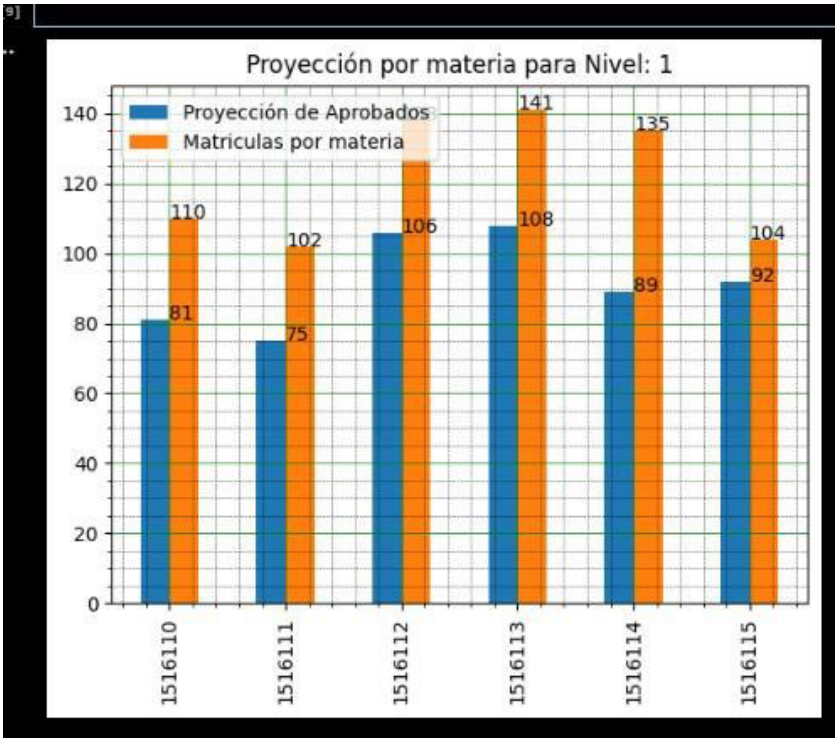
Python

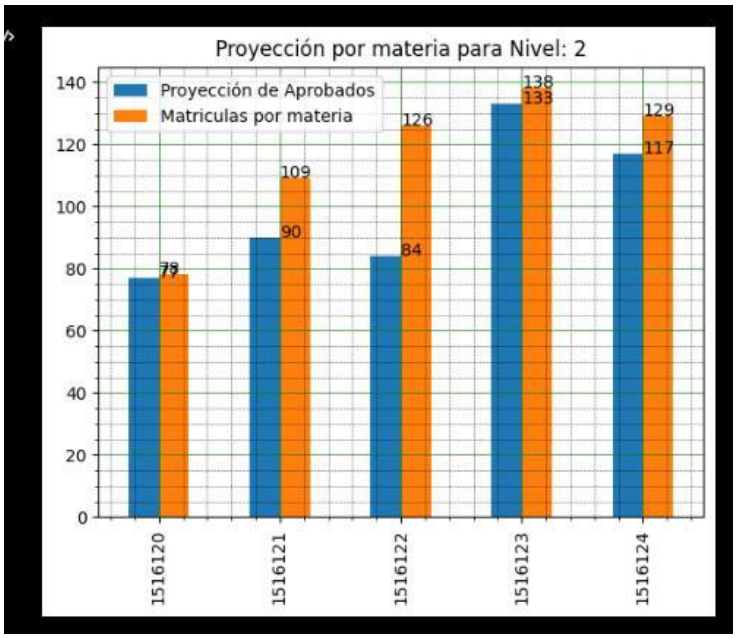


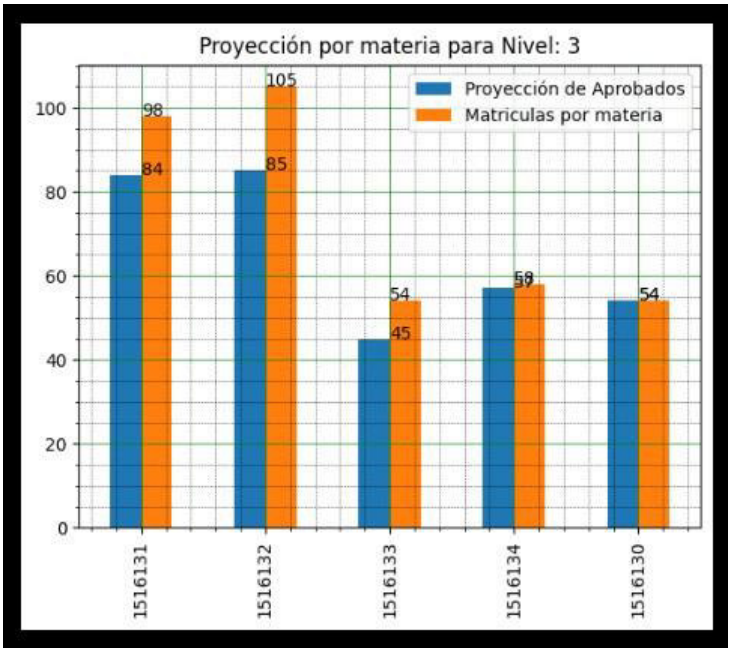
```
ArchiDirec=os.getcwd()+"\\data\\" # Ruta de datos
ArchivoTest=ArchiDirec + "datatest.xlsx"
DataTest=pd.read_excel(ArchivoTest)

# Calculo de promedio de eficiencia de estudiantes de nivel 2
# Se usa el nivel 2 ya que estos estudiantes son los que anteriormente estaban en nivel 1
# en el periodo analizado, y es una buena referencia de como se comportan los estudiantes
# en periodos contemporaneos


# Obtener los valores filtrados correspondiente al Nivel 2
CalEfi=(DataTest.apply(lambda x: x['NIVEL_MAT'] == 2, axis=1))
CalEfiN2 = DataTest[CalEfi]['EfiEstud'].tolist()

# Calcular el promedio de la eficiencia de los estudiantes de nivel 2
CalEfiP2=statistics.mean(CalEfiN2)

# Actualizar valores nulos de eficiencia del estudiante por el promedio previamente calculado
DataTest['EfiEstud'].fillna(CalEfiP2, inplace=True)

#DataTest.info()

# Calculo de promedio de eficiencia de docentes en un solo paso
CalEfiPDoc  = statistics.mean(DataTest[DataTest['EfiDocen'].notna()]['EfiDocen'].tolist())

# print (CalEfiPDoc)

# Actualizar valores nulos de eficiencia docente por el promedio previamente calculado
DataTest['EfiDocen'].fillna(CalEfiPDoc, inplace=True)
#DataTest.info()

# Calculo de promedio de eficiencia de materias
CalEfiPMat  = statistics.mean(DataTest[DataTest['EfiMater'].notna()]['EfiMater'].tolist())
```

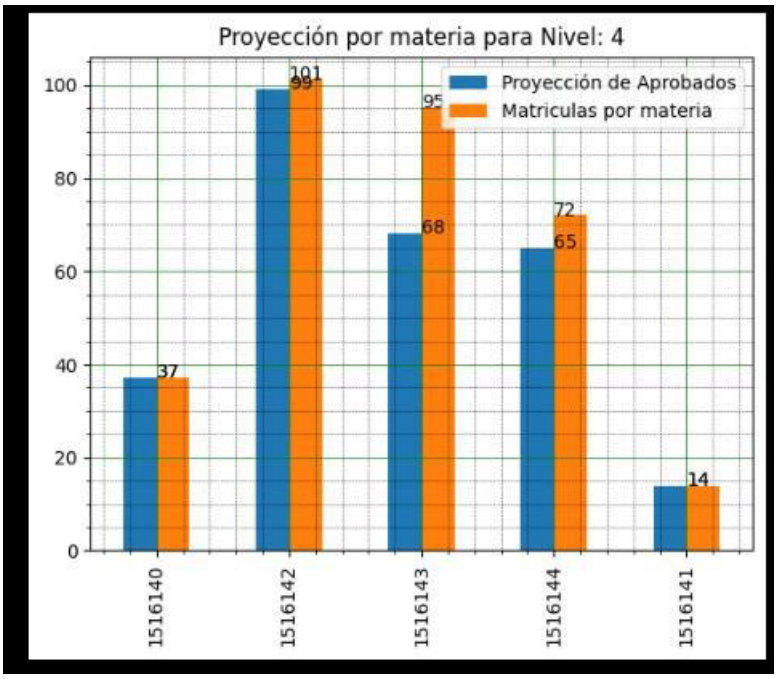
```
def add_value_label(x_list,y_list):
    for i in range(1, len(x_list)+1):
        plt.text(i-1.0,y_list[i-1],y_list[i-1])

# Se carga los datos para hacer las predicciones
ArchiDirec=os.getcwd()+"\\data\\" # Ruta de datos
ArchivoTest=ArchiDirec + "datatest.xlsx"
DataTest=pd.read_excel(ArchivoTest)

# Calculo de promedio de eficiencia de estudiantes de nivel 2
# Se usa el nivel 2 ya que estos estudiantes son los que anteriormente estaban en nivel 1
# en el periodo analizado, y es una buena referencia de como se comportan los estudiantes
# en periodos contemporaneos


# Obtener los valores filtrados correspondiente al Nivel 2
CalEfi=(DataTest.apply(lambda x: x['NIVEL_MAT'] == 2, axis=1))
CalEfiN2 = DataTest[CalEfi]['EfiEstud'].tolist()

# Calcular el promedio de la eficiencia de los estudiantes de nivel 2
CalEfiP2=statistics.mean(CalEfiN2)

# Actualizar valores nulos de eficiencia del estudiante por el promedio previamente calculado
DataTest['EfiEstud'].fillna(CalEfiP2, inplace=True)

#DataTest.info()

# Calculo de promedio de eficiencia de docentes en un solo paso
CalEfiPDoc  = statistics.mean(DataTest[DataTest['EfiDocen'].notna()]['EfiDocen'].tolist())

# print (CalEfiPDoc)

# Actualizar valores nulos de eficiencia docente por el promedio previamente calculado
DataTest['EfiDocen'].fillna(CalEfiPDoc, inplace=True)
#DataTest.info()
```

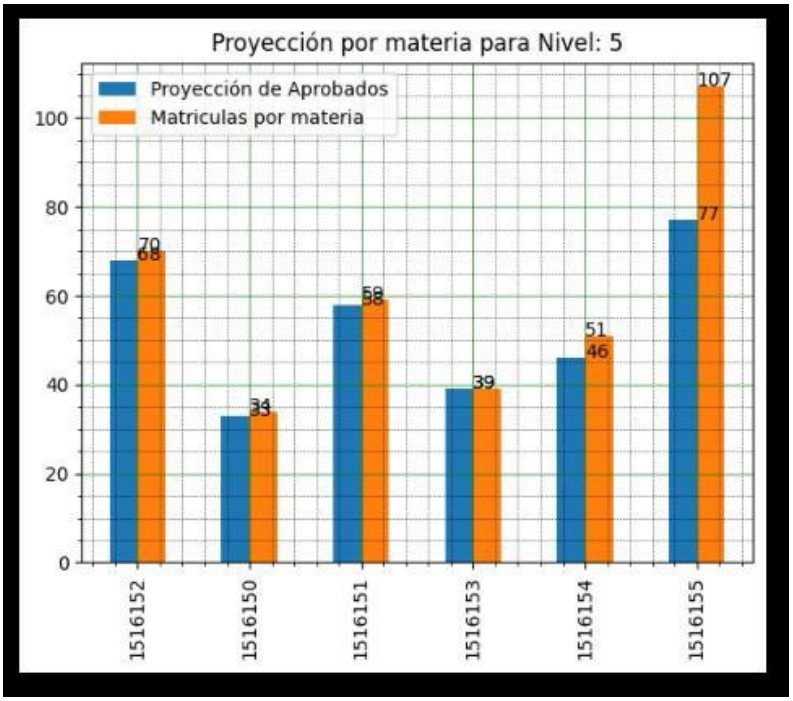
```
# Actualizar valores nulos de eficiencia docente por el promedio previamente calculado
DataTest['EfiMater'].fillna(CalEfiPMat, inplace=True)
#DataTest.info()

# Procedimiento de limpieza de datos en caso de que existan valores nulos para Asistencia o Promedio
DataTest.drop(DataTest[(DataTest["ASISTENCIA 1P"].isnull() )].index, inplace=True)
DataTest.drop(DataTest[(DataTest["PROMEDIO P1"].isnull() )].index, inplace=True)

xTest= DataTest.copy().iloc[:,1:]
DataTest["ESTADOAP"] = DataArbolDecis.predict(xTest)
DataTest.to_csv(ArchiDirec + "proyeccion.csv",index=False)
#DataTest.to_excel(ArchiDirec + "proyeccion.xlsx",index=False)


for i in range(10):
  DataPrediNiv=DataTest.loc[DataTest["NIVEL_MAT"] == (i+1)]
  DataMaterNiv=np.array(DataPrediNiv["COD_MATERIA"].unique()) # se busca las filas unicas en el campo "COD_MAT"
  c2=[] # Almacena la cantidad de estudiantes que tienen la predicción de Aprobado
  c3=[] # Almacena la cantidad de estudiantes registrados en la materia
  for DataGrafi01 in DataMaterNiv: # Se preparan los datos que se mostrarán en la gráfica
    c2.append(DataPrediNiv.apply(lambda x: x['ESTADOAP'] == 1 and x["COD_MATERIA"] == DataGrafi01, axis=1).sum() )
    c3.append(DataPrediNiv.apply(lambda x: x["COD_MATERIA"] == DataGrafi01, axis=1).sum() )

  DataGrafiTest=pd.DataFrame({'Proyección de Aprobados':c2, "Matriculas por materia":c3}, index=DataMaterNiv)
  DataGrafiTest.plot(kind='bar', title=f'Proyección por materia para Nivel: {i+1}')

  add_value_label(DataMaterNiv,c2)
  add_value_label(DataMaterNiv,c3)

  plt.minorticks_on() # Para activar las grillas en el gráfico
  plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
  plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
  plt.show()
```

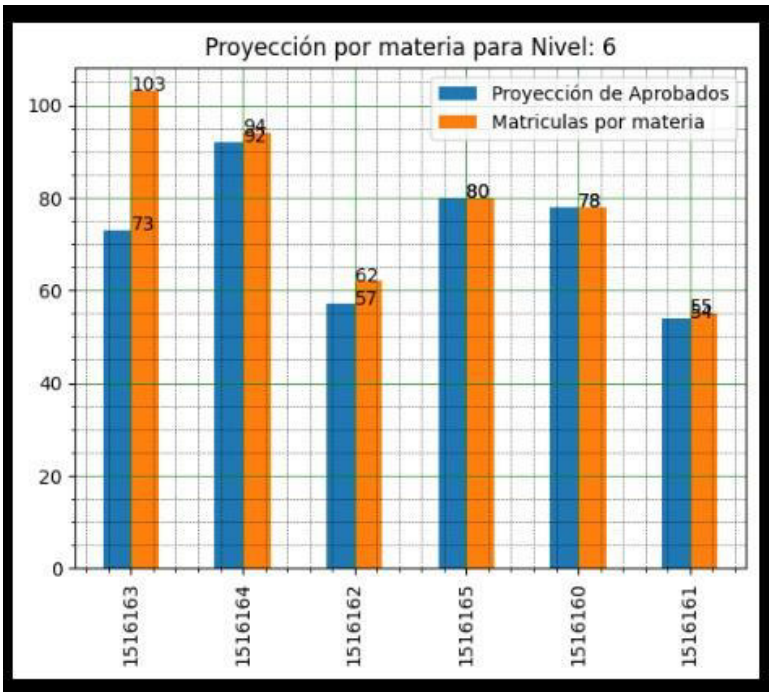


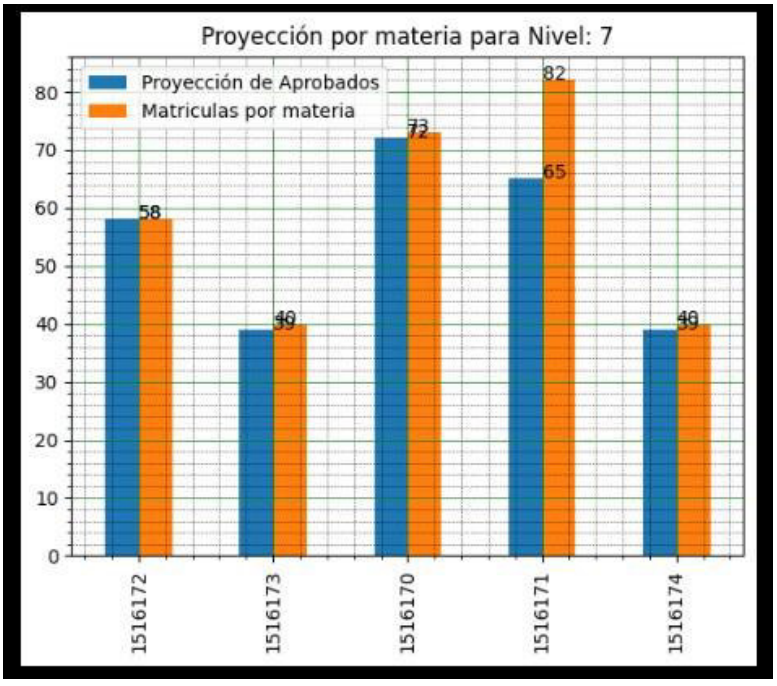

Proyección por materia para Nivel: 2
Proyección de Aprobados
Matriculas por materia
140
120
100
80
60
40
20
0
78
78
90
109
84
126
133
138
117
129
1516120
1516121
1516122
1516123
1516124

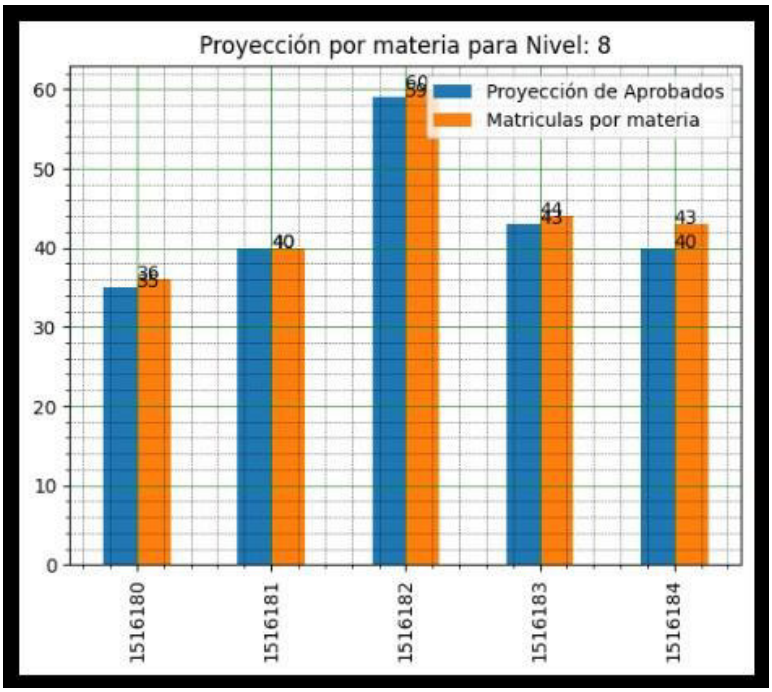

Proyección por materia para Nivel: 3
Proyección de Aprobados
Matriculas por materia
100
80
60
40
20
0
84
98
85
105
45
54
57
58
54
1516131
1516132
1516133
1516134
1516130

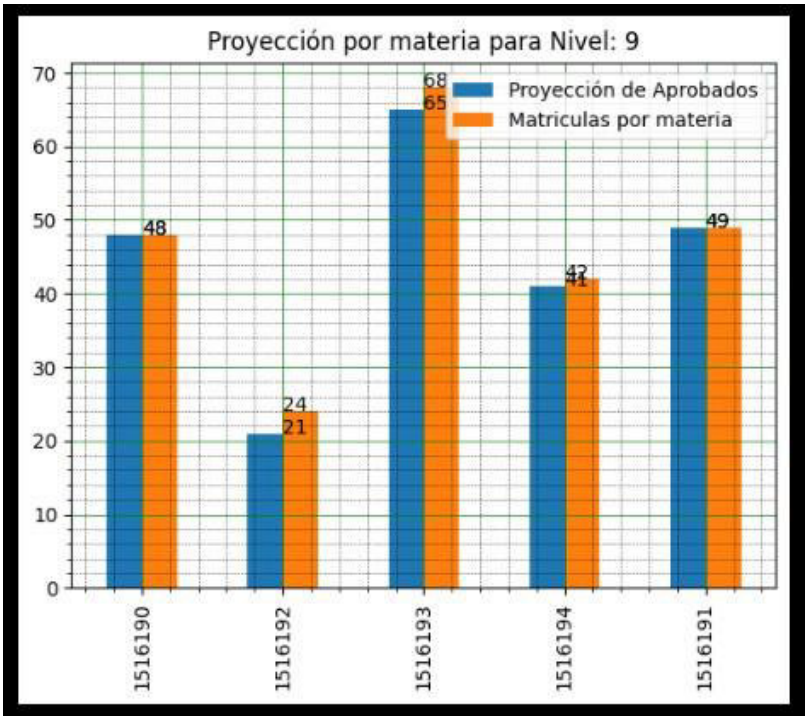

Proyección por materia para Nivel: 4
Proyección de Aprobados
Matriculas por materia
100
80
60
40
20
0
37
101
68
72
65
14
1516140
1516142
1516143
1516144
1516141

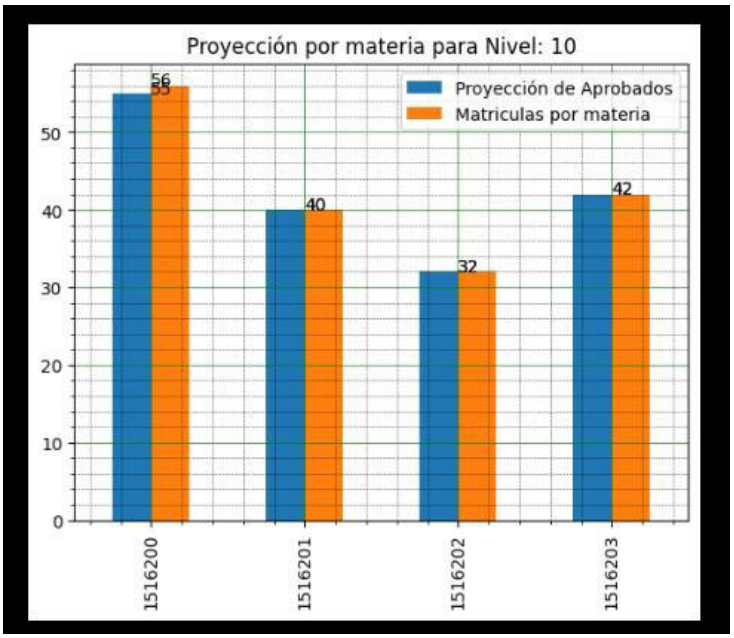

Proyección por materia para Nivel: 5
Proyección de Aprobados
Matriculas por materia
100
80
60
40
20
0
70
68
34
58
39
51
46
77
107
1516152
1516150
1516151
1516153
1516154
1516155


Proyección por materia para Nivel: 6
Proyección de Aprobados
Matriculas por materia
103
73
94
92
62
57
80
78
55
54
0
20
40
60
80
100
1516163
1516164
1516162
1516165
1516160
1516161


Proyección por materia para Nivel: 7
Proyección de Aprobados
Matriculas por materia
58
40
73
72
65
82
40
0
10
20
30
40
50
60
70
80
1516172
1516173
1516170
1516171
1516174


Proyección por materia para Nivel: 8
Proyección de Aprobados
Matriculas por materia
60
59
44
43
43
40
40
36
35
0
10
20
30
40
50
60
1516180
1516181
1516182
1516183
1516184


Proyección por materia para Nivel: 9
Proyección de Aprobados
Matriculas por materia
68
65
49
48
48
42
41
24
21
0
10
20
30
40
50
60
70
1516190
1516192
1516193
1516194
1516191



Modelado Random Forest

```
# Importar las bibliotecas necesarias
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

ArchiDirec=os.getcwd()+"\\data\\" # Ruta de datos=carpeta actual\\data\\
Archivo=ArchiDirec + "datatrain.xlsx" # Archivo con la información en formato xlsx
DataCalif = pd.read_excel(ArchiDirec + "datatrain.xlsx")
# Cargar los datos de prueba desde un archivo CSV
data_test = pd.read_excel(ArchiDirec + "datatest.xlsx")

# El DataFrame usado es lo suficientemente grande y contiene suficientes opciones de aprobado o reprobado
# Localizamos los duplicados segun los campos que deben dar datos únicos, escogiendo la primera aparición de ser el caso
DataDuplicada = DataCalif.duplicated(subset = ["PERIODO","IDENTIFICACION","NIVEL_MAT","COD_MATERIA","CCDOCEN"], keep = 'first')

# Buscamos y eliminamos las filas que tengan valores nulos en el campo "ASISTENCIA" y "PROM FINAL", y "PROM P1" ya que se necesita la información completa en estos campos
DataCalif.drop(DataCalif[(DataCalif["ASISTENCIA 1P"].isnull() )].index, inplace=True)
DataCalif.drop(DataCalif[(DataCalif["PROMEDIO P1"].isnull() )].index, inplace=True)


# Se escoje las columnas que se van a utilizar para el analisis predictivo
# (variables independientes) y se asignan a la variable "x"
x= DataCalif.copy().iloc[:,1:10]
```

```python
# Se escoje las columnas que se van a utilizar para el analisis predictivo
# (variables independientes) y se asignan a la variable "x"
x= DataCalif.copy().iloc[:,1:10]

# Se escoje la columna de resultados (variable dependiente)
# se asignan a la variable "y"
y= DataCalif["ESTADOAP"]


# Se divide la muestra en datos de entrenamiento y de prueba#, random_state=42)
# el atributo  random_state= es para que siempre se escoja los mismos datos
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.75)


# Crear un modelo Random Forest
model = RandomForestClassifier(n_estimators=100, random_state=42)  # Puedes ajustar el número de árboles (n_estimators) según tus necesidades

# Entrenar el modelo con los datos de entrenamiento
model.fit(X_train, y_train)

# Realizar predicciones en el conjunto de prueba
y_pred = model.predict(X_test)

# Calcular la precisión del modelo
accuracy = accuracy_score(y_test, y_pred)
print(f"Precisión del modelo Random Forest: {accuracy:.2f}")

# Mostrar un informe de clasificación
report = classification_report(y_test, y_pred)
print("Informe de clasificación:\n", report)
```

```
[12]
...  Precisión del modelo Random Forest: 0.88
     Informe de clasificación:
                   precision    recall  f1-score   support

                0       0.73      0.60      0.66      3178
                1       0.91      0.95      0.93     14064

         accuracy                           0.88     17242
        macro avg       0.82      0.77      0.79     17242
     weighted avg       0.88      0.88      0.88     17242
```

```python
# Importar las bibliotecas necesarias
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, classification_report


#X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Escalar las características (opcional pero generalmente recomendado)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Crear un modelo de Regresión Logística
model = LogisticRegression()
```

```python
#X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Escalar las características (opcional pero generalmente recomendado)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Crear un modelo de Regresión Logística
model = LogisticRegression()

# Entrenar el modelo
model.fit(X_train, y_train)

# Realizar predicciones en el conjunto de prueba
y_pred = model.predict(X_test)

# Calcular la precisión del modelo
accuracy = accuracy_score(y_test, y_pred)
print(f"Precisión del modelo de Regresión Logística: {accuracy:.2f}")

# Mostrar un informe de clasificación
report = classification_report(y_test, y_pred)
print("Informe de clasificación:\n", report)
```

```
[14]
...  Precisión del modelo de Regresión Logística: 0.89
     Informe de clasificación:
                   precision    recall  f1-score   support

                0       0.76      0.55      0.64      3178
                1       0.90      0.96      0.93     14064

         accuracy                           0.89     17242
        macro avg       0.83      0.76      0.79     17242
     weighted avg       0.88      0.89      0.88     17242
```

BIBLIOGRAFÍA

Abu Kausar, M. a. (1532-1540). A Study of Performance and Comparison of NoSQL Databases: MongoDB, Cassandra, and Redis Using YCSB. *Indian Journal of Science and Technology*, 2022.

Anusha, K. a. (2021). Comparative Study of MongoDB vs Cassandra in big data analytics. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1831-1835.

Auccapuri, A. A. (2021). Planificación curricular en la enseñanza universitaria y desempeño profesional de egresados en educación. *Ciencia Latina Revista Científica Multidisciplinar*, 2563-2589.

AWS. (2023). *Python* . Obtenido de https://aws.amazon.com/es/what-is/python/

Ayinuer, N. a. (2022). Design and Research of Unstructured Data Knowledge Graph Toolbased on Neo4j Graph Database. *2022 11th International Conference on Communications, Circuits and Systems (ICCCAS)*, 296-300.

Balzer, W. K. (2020). *Lean higher education: Increasing the value and performance of university processes.* CRC Press.

Chango Gavilánez, D. Y. (2016). *Bases de datos no relacionales: Utilización de Mongo DB como base de datos no relacional empleando formato GEO JSON.* Bachelor's thesis.

Chauhan, A. (2019). A review on various aspects of MongoDb databases. *Int. J. Eng. Res. Sci. Technol*, 90-92.

Contreras, L. E. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación universitaria*, 233-246.

DataScientest. (2023). Obtenido de https://datascientest.com/es/scikit-learn-decubre-la-biblioteca-python

DataScientest. (2023). *pandas*. Obtenido de https://datascientest.com/es/pandas-python

Del Pozo Puñal, E. (2019). *Diseño y desarrollo de un sistema de ficheros distribuido y paralelo basado en Apache Cassandra.*

DOCKER. (2023). *qué es Docker*. Obtenido de https://www.docker.com/

Gupta, S. &. (2020). Academic Staff planning, allocation and optimization using Genetic Algorithm under the framework of Fuzzy Goal Programming. *Procedia Computer Science*, 900-905.

Hadida, S. &. (2020). La agilidad en las organizaciones: Trabajo comparativo entre metodología a giles y de cascada en un contexto de ambigüedad y transformación digital. *Serie Documentos de Trabajo*, 756.

Haris, L. (2018). Risk Assessment on Information Asset an academic Application Using ISO 27001. *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 1-4.

Holguín Quimis, R. J. (2020). Estudio de factibilidad de un sistema de control de acceso con tecnología RFID para la unidad de bienestar estudiantil de la Universidad Estatal del Sur de Manabí. *Tesis de Licenciatura*.

Huamantumba, E. J. (2020). Gestión de simplificación administrativa en el desarrollo de las universidades públicas. *Hacedor-AIAPÆC*, 69-82.

Hunter, J. (2023). *Matplotlib*. Obtenido de https://matplotlib.org/

Link, S. (2020). Neo4j keys. *Conceptual Modeling: 39th International Conference, ER 2020, Vienna, Austria, November 3--6, 2020, Proceedings 39* (págs. 19-33). Springer.

McCaffery, P. (2018). *The higher education manager's handbook: effective leadership and management in universities and colleges.* Routledge.

Microsoft. (2023). *qué es power bi*. Obtenido de https://powerbi.microsoft.com/es-es/what-is-power-bi/

Mihiranga, N. (2022). *Power BI Data Modeling: Build Interactive Visualizations, Learn DAX, Power Query, and Develop BI Models (English Edition.* BPB Publications.

Narváez, M. E. (2020). Análisis de Desempeño entre MONGODB y COUCHDB utilizando Norma ISO/IEC 25000. *Revista Perspectivas*, 13-20.

NUMPY. (2023). *Qué es Numpy*. Obtenido de https://numpy.org/

Panchi Arias, M. P. (2021). La auditoría interna como herramienta de control y seguimiento de la gestión en las universidades. *Revista Universidad y Sociedad*, 333-341.

Piccardi, M. L. (2021). Del BIG DATA al FAST DATA: enfoques modernos de streaming de datos para el procesamiento de datos masivos en tiempo real. *Difusiones*, 38-58.

República del Ecuador. (2004). *Ley de Gestión Ambiental.* Quito: Ministerio del Ambiente, Agua y Transición Ecológica.

República del Ecuador. (2015). *Reglamento General a La Ley Orgánica de Educación Intercultural .* Quito: Ministerio de Eduación.

República del Ecuador. (2018). *Ley Orgánica de Eduación Superior, LOES.* República del Ecuador. Quito: Registro Oficial Suplemento 298 de 12-oct.-2010.

República del Ecuador. (2019). *Reglamento de Distribuición Recursos Instituciones Educación Superior.* Quito: Consejo de Educación Superior.

República del Ecuador. (2021). *Ley Orgánica de Protección de Datos Personales.* Quito: Ministerio de Telecomunicaciones y de la Sociedad de la Información.

República del Ecuador. (2022). *Política Para La Transformación Digital Ecuador 2022-2025.* Quito: Ministerio de Telecominicaciones y Sociedad de la Información .

Reyes-González, N. M.-B.-M. (2022). Planificación y gestión del tiempo académico de estudiantes universitarios. *Formación universitaria*, 57-72.

Tonysé de la Rosa, M. (2021). Automatización de un sistema de gestión de seguridad de la información basado en la Norma ISO/IEC 27001. *Revista Universidad y Sociedad*, 495-506.

Vargas-Larraguível, P. (2021). *Factores de impacto en la información emprendedora en estudiantes de educación superior.* CETYS Universidad.

Wahid, A. a. (2019). Cassandra—A distributed database system: An overview. *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 1*, 519-526.