



**Escuela Superior
de Ingeniería y Tecnología**
Universidad de La Laguna

Trabajo de Fin de Grado

Business Intelligence y Machine Learning para la explotación de datos educativos

*Business Intelligence and Machine Learning for the
exploitation of educational data*

Fabrizio Daniell Perilli Martín

La Laguna, 11 de julio de 2024

D. **Pedro Antonio Toledo Delgado**, con N.I.F. 45.725.874-B profesor contratado Doctor, adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutor

C E R T I F I C A (N)

Que la presente memoria titulada:

"Business Intelligence y Machine Learning para la explotación de datos educativos "

ha sido realizada bajo su dirección por D. **Fabrizio Daniell Perilli Martín**, con N.I.F. 42.291.244-V

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 11 de julio de 2024

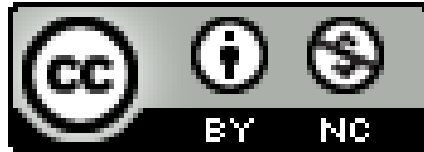
Agradecimientos

Quiero expresar mi agradecimiento a todas las personas que han contribuido en el desarrollo de mis estudios. En primer lugar, me gustaría agradecer a mi tutor, Pedro Toledo Delgado, quien siempre demostró su apoyo y asesoría durante todo el desarrollo del proyecto. Su colaboración ha sido un pilar clave para la finalización de este trabajo.

También quiero agradecer a mis amigos y compañeros de clase. He conocido a buenas personas en el camino que me han aportado cosas positivas, considero que han sido fundamentales en mi desarrollo tanto personal como profesional.

Finalmente, quiero agradecer a mi familia, especialmente a mis padres y hermanos. Son las personas que han confiado en mí desde que comencé el grado, siempre han estado presente para apoyarme en todo momento. En particular, quiero agradecer a mi madre, es la persona más importante para mí y un ejemplo a seguir. Gracias a ella he llegado hasta aquí.

Licencia



© Esta obra está bajo una licencia de Creative Commons Reconocimiento-
NoComercial 4.0 Internacional.

Resumen

El abandono académico es una problemática importante que produce consecuencias considerables en nuestra sociedad, como el incremento del desempleo o la desigualdad social y económica. Esta situación afecta el desarrollo personal y profesional de las personas y supone un gasto adicional en recursos públicos.

Dada la naturaleza de las universidades, estas entidades acumulan una gran cantidad de datos, principalmente administrativos, sobre los estudiantes. Este recurso es valioso para identificar áreas de oportunidad y riesgo. Sin embargo, la falta de herramientas adecuadas impide que la información disponible se utilice de manera efectiva.

Para abordar esta problemática, se ha desarrollado una herramienta analítica. Este software utiliza métodos de Machine Learning y Business Intelligence para examinar, estructurar y convertir datos sin procesar en información valiosa. La aplicación proporciona cuadros de mando interactivos para visualizar el progreso académico de los estudiantes y, además, cuenta con un modelo de aprendizaje automático que permite predecir la probabilidad de que los alumnos abandonen sus estudios.

Los resultados obtenidos del proyecto podrían ofrecer una mejora significativa con respecto a la visualización, comprensión y análisis de los datos, además de poder aportar un valor representativo para la toma de decisiones por parte de los interesados.

Palabras clave: Inteligencia empresarial, Análisis de datos, Abandono académico universitario, Análisis del aprendizaje.

Abstract

Academic dropout is a major problem that has considerable consequences in our society, such as increased unemployment or social and economic inequality. This situation affects the personal and professional development of individuals and represents an additional expense in public resources.

Given the nature of universities, these entities accumulate a large amount of data, mainly administrative, on students. This resource is valuable for identifying areas of opportunity and risk. However, the lack of adequate tools prevents the available information from being used effectively.

To address this issue, an analytical tool has been developed. This software uses Machine Learning and Business Intelligence methods to examine, structure, and convert raw data into valuable information. The application provides interactive dashboards to visualize students academic progress and also includes a machine learning model to predict the probability of students dropping out of school.

The results obtained from the project could offer significant improvement with respect to visualization, understanding, and analysis of the data. Additionally, they could provide substantial value for decision-making by stakeholders.

Keywords: Business Intelligence, Data Analysis, University academic dropout, Learning analytics.

Índice general

1. Introducción	1
1.1. Contexto y justificación	1
1.2. Objetivos generales	3
1.3. Objetivos específicos	3
2. Marco teórico y antecedentes	4
2.1. Business Intelligence	4
2.1.1. Ciclo de vida	4
2.1.2. Aplicación en el sector académico	5
2.2. Machine Learning	6
2.3. Estado del arte	6
3. Tecnologías	8
3.1. Tecnologías utilizadas	8
3.2. Análisis de la herramienta BI	10
3.2.1. Funcionamiento	10
3.2.2. Escalabilidad	11
3.2.3. Despliegue	12
4. Desarrollo	13

4.1. Generación de datos	13
4.1.1. Diseño de base de datos	13
4.1.2. Datos sintéticos	16
4.2. Diseño de visualización de datos	17
4.2.1. Perfil alumnado	18
4.2.2. Perfil docente	19
4.2.3. Perfil gestor	20
4.2.4. Validación del diseño	21
4.3. Diseño y arquitectura de software	21
4.4. Modelo de aprendizaje automático	23
4.4.1. Recogida de datos	23
4.4.2. Procesamiento de los datos	23
4.4.3. Selección del algoritmo	26
4.4.4. Entrenamiento y evaluación de los modelos	28
5. Resultados	30
5.1. Visualización y funcionalidades	30
5.1.1. Perfil alumnado	30
5.1.2. Perfil docente	36
5.1.3. Perfil gestor	40
5.2. Resultados del modelo predictivo	47
5.3. Divulgación de la herramienta	50
6. Conclusiones y líneas futuras	51
6.0.1. Conclusiones	51
6.0.2. Líneas futuras	52

7. Summary and Conclusions	53
8. Presupuesto	54

Índice de Figuras

4.1. Esquema de la base datos inicial (Origen proyecto ABACU).	14
4.2. Esquema de base de datos utilizado.	15
4.3. Primera versión mockup. Perfil alumnado.	18
4.4. Primera versión mockup. Perfil docente.	19
4.5. Primera versión mockup. Perfil gestor.	20
4.6. Estructura de directorios del proyecto.	22
5.1. Perfil alumnado. Pestaña expediente académico personal.	31
5.2. Gráfica. Evolución del progreso académico.	32
5.3. Gráfica. Evolución calificaciones cualitativas.	32
5.4. Gráfica. Calificaciones cuantitativas.	32
5.5. Gráfica. Tasa de rendimiento académico.	33
5.6. Perfil alumnado. Pestaña rendimiento académico general.	33
5.7. Gráfica. Relación nota media y asignaturas superadas.	34
5.8. Gráfica. Calificaciones cualitativas general.	34
5.9. Gráfica. Relación calificaciones del alumno con nota media general.	35
5.10 Perfil alumnado. Pestaña recomendaciones.	35
5.11 Perfil docente. Pestaña rendimiento académico personal.	36
5.12 Gráfica. Evolución alumnos de nuevo ingreso y repetidores	37
5.13 Gráfica. Evolución alumnos matriculados por género	37

5.14	Gráfica. Evolución calificaciones cualitativas	37
5.15	Gráfica. Evolución nota media por asignatura	37
5.16	Perfil docente. Pestaña rendimiento académico general.	38
5.17	Gráfica. Calificaciones cualitativas de la titulación.	38
5.18	Gráfica. Nota media por asignaturas de la titulación.	39
5.19	Perfil docente. Pestaña recomendaciones.	39
5.20	Perfil gestor. Pestaña indicadores.	40
5.21	Gráfica. Alumnos de nuevo ingreso por género y titulación.	41
5.22	Gráfica. Alumnos de nuevo ingreso por nacionalidad y titulación.	41
5.23	Gráfica. Alumnos egresados por género y titulación.	41
5.24	Gráfica. Alumnos egresados por nacionalidad y titulación.	42
5.25	Perfil gestor. Pestaña resultados.	42
5.26	Gráfica. Duración media de los estudios con respecto a la nota media.	43
5.27	Gráfica. Evolución nota de acceso por titulación.	43
5.28	Perfil gestor. Pestaña riesgo de abandono.	44
5.29	Gráfica. Tasa de abandono por titulación.	44
5.30	Gráfica. Tasa de graduación por titulación.	45
5.31	Perfil gestor. Pestaña recomendaciones.	45
5.32	Ejemplo herramientas de gráficos.	46
5.33	Tabla de datos.	46
5.34	Matriz de confusión.	47
5.35	Resultado de probabilidad de abandono. Perfil alumnado	49

Índice de Tablas

3.1. Tabla comparativa de herramientas BI.	8
5.1. Resultados de los modelos	48
5.2. Matrices de confusión de los modelos	49
8.1. Resumen del Presupuesto Desglosado	54

Capítulo 1

Introducción

1.1. Contexto y justificación

Las universidades representan un pilar fundamental en nuestra sociedad, ya que promueven el desarrollo a través de la educación y la investigación. Una de las problemáticas que enfrentan las instituciones, y que ha persistido en el tiempo, es el abandono académico. Este representa un factor de riesgo significativo para los estudiantes y las universidades. A nivel nacional, según un estudio realizado por el Ministerio de Universidades titulado *“Análisis del abandono de los estudiantes de grado en las universidades presenciales en España”*, se indica que, según los datos obtenidos del Sistema Integrado de Información Universitaria (SIIU), el 20,8 % de los estudiantes de nuevo ingreso en el curso 2015-2016 abandonaron el grado [13].

La deserción académica tiene un impacto significativo en la sociedad y en la economía. Los estudiantes pueden enfrentarse a pocas oportunidades laborales, ingresos más bajos y menos estabilidad económica. En el caso de las universidades, una alta tasa de abandono universitario puede resultar en una menor eficiencia y en un aumento en el gasto público para la educación. Una investigación estima que el fracaso escolar supuso un sobrecoste de más de 4.500 millones de euros en España en el año 2022. De esta cantidad, 604,9 millones de euros correspondieron a los estudiantes de grado [3].

La investigación sobre el abandono académico ha sido abordada desde diferentes perspectivas teóricas y metodológicas. Un análisis realizado por Vincent Tinto (1993) [25] destaca que la persistencia de los estudiantes está influenciada por su grado de integración en el entorno académico y social de la universidad. Los estudiantes que se sienten desconectados de su comunidad educativa tienen más probabilidades de abandonar sus estudios. Otra investigación relevante es la de Bean y Metzner (1985) [1], quienes desarrollaron un modelo de abandono para estudiantes no tradicionales. Su estudio sugiere que los factores externos a la institución, como las obligaciones laborales y familiares, tienen un impacto significativo en la decisión de los estudiantes de abandonar sus estudios.

La retención de estudiantes es un asunto complejo que se ve afectado por factores personales, institucionales y sociales. El rendimiento académico, las dificultades de aprendizaje, la falta de motivación, la insatisfacción con la carrera elegida, problemas personales como la ansiedad o la depresión, y la falta de estrategias de aprendizaje son factores personales que contribuyen al abandono académico. Los factores institucionales, como la calidad de la docencia, la oferta de servicios de apoyo académico y la existencia de ayudas económicas, también son importantes. Finalmente, los factores sociales que influyen en el abandono incluyen el apoyo familiar y social, la situación socioeconómica de los estudiantes, las expectativas del entorno y las condiciones del mercado laboral.

En este contexto, el análisis de datos es clave para detectar debilidades estudiantiles. El uso de técnicas de minería de datos y aprendizaje automático ha demostrado ser eficaz en la predicción del riesgo de abandono. Por ejemplo, un estudio de Marbouti, Diefes-Dux y Madhavan (2016) [10] combinó algoritmos de clasificación y regresión para predecir el abandono académico. De manera similar, Kovacic (2010) [7] utilizó técnicas de análisis discriminante y regresión logística para identificar las variables más influyentes en el abandono de estudiantes en un entorno de educación a distancia. Los resultados de estos estudios muestran que los factores demográficos, el rendimiento académico y la interacción con los recursos de la universidad son predictores significativos del abandono.

En el año 2020, se aprobó el desarrollo de un proyecto financiado por el Ministerio de Ciencia e Innovación de España, titulado *“Análisis de los factores explicativos del abandono de los estudios universitarios y acciones estratégicas para su mejora y prevención”* (ABACU) [28]. Coordinado por la Universidad de La Laguna, involucra también a las universidades de Castilla-La Mancha, Santiago de Compostela, Zaragoza y Huelva. El proyecto se centra en analizar los factores que influyen en el abandono, identificar a estudiantes en riesgo, diseñar estrategias para mejorar los índices de deserción y visibilizar los avances de la investigación.

El Trabajo de Fin de Grado presentado aporta un valor añadido y actúa como complemento a las investigaciones realizadas en el proyecto mencionado anteriormente. Se centra en utilizar la información de las universidades de manera eficiente para aprovechar el potencial de los recursos disponibles. También se consideran aspectos de usabilidad y accesibilidad para garantizar que los usuarios puedan aprovechar al máximo las funcionalidades ofrecidas.

Las representaciones gráficas son fundamentales para identificar patrones y tendencias, por lo que se ha desarrollado una herramienta que incluye cuadros de mando interactivos y técnicas de aprendizaje automático que permiten predecir la probabilidad de abandono académico de cada alumno, basándose en datos históricos y actuales. Se busca mejorar la representación de los datos obtenidos, permitiendo una visión integral de los resultados académicos para facilitar la ejecución de un plan estratégico por parte de los responsables.

1.2. Objetivos generales

Desarrollar una aplicación web analítica e interactiva, de libre uso y distribución, utilizando herramientas de Business Intelligence y técnicas de Machine Learning para mejorar la comprensión, visualización y análisis del rendimiento estudiantil, así como para facilitar la toma de decisiones estratégicas informadas.

1.3. Objetivos específicos

1. Realizar un análisis comparativo de las herramientas de Business Intelligence, evaluando sus características para lograr identificar la mejor opción para el desarrollo de la aplicación.
2. Analizar las representaciones visuales y los casos de uso del sistema.
3. Generar de datos sintéticos para visualizar los resultados del proyecto.
4. Diseñar e implementar cuadros de mando interactivos que faciliten la visualización dinámica de los resultados académicos.
5. Desarrollar, entrenar e integrar un modelo predictivo que estime la probabilidad de abandono académico de cada estudiante.

Capítulo 2

Marco teórico y antecedentes

2.1. Business Intelligence

El término Business Intelligence (BI) o Inteligencia de Negocio fue acuñado por el investigador de IBM, Hans Peter Luhn, en 1958, quien lo definió como la *“Capacidad de comprender las relaciones entre los hechos para guiar las acciones hacia una meta deseada”* [9]. Este campo incluye un conjunto de técnicas y herramientas destinadas a recopilar, almacenar, analizar y permitir el acceso a los datos. Su objetivo es brindar información útil y precisa que ayude a los usuarios a tomar decisiones estratégicas.

2.1.1. Ciclo de vida

La implementación de un sistema BI sigue un proceso cíclico y sistemático que garantiza que el sistema se alinee con los objetivos, cumpla con las necesidades de los usuarios y se adapte a los cambios:

- **Recopilación de requisitos:** Comienza con la determinación del objetivo, examinando los requisitos y el propósito del análisis.
- **Preparación de los datos:** El siguiente paso es diseñar el modelo lógico y el esquema físico para realizar el análisis de los datos. Los datos deben organizarse y limpiarse adecuadamente.
- **Informe de análisis:** Los datos agregados y consolidados en un único lugar se convierten en informes visualmente atractivos. Los cuadros de mando interactivos permiten a las partes interesadas comunicar y profundizar en las tendencias y patrones.

- **Tomar acción:** En esta fase, se aprovecha el poder de los sistemas BI. Los responsables de la toma de decisiones pueden medir y supervisar los indicadores clave de rendimiento para alinear la estrategia con los objetivos.

La última etapa del ciclo BI es la iteración del proceso. Se evalúa el éxito de las decisiones y se realizan ajustes para mejorar los resultados. Esto puede implicar una reevaluación completa de los requisitos, volviendo al primer paso y creando un proceso iterativo [6].

2.1.2. Aplicación en el sector académico

En el sector académico, se utilizan herramientas de BI para mejorar la gestión y el rendimiento educativo mediante el análisis de grandes volúmenes de datos sobre los involucrados en el sector académico [21]. Los sistemas en las universidades permiten:

- **Monitoreo del rendimiento estudiantil:** Los cuadros de mando y los informes detallados ayudan a identificar a los estudiantes en riesgo de abandono, permitiendo la intervención temprana y personalizada.
- **Mejora de la toma de decisiones:** Los administradores y educadores pueden tomar decisiones informadas sobre programas de estudio y políticas institucionales basándose en datos precisos y análisis detallados.
- **Incremento de la retención y el éxito académico:** Al identificar patrones y tendencias en el rendimiento académico, las instituciones pueden implementar estrategias para mejorar la retención de estudiantes y aumentar las tasas de graduación.

2.2. Machine Learning

El Machine Learning (aprendizaje automático en español) es una rama de la inteligencia artificial, transforma la manera en que las computadoras procesan la información, permitiéndoles aprender de los datos, identificar patrones, hacer predicciones y mejorar su desempeño en diversas aplicaciones. Este enfoque está revolucionando sectores como finanzas, tecnología, salud y muchos otros.

El aprendizaje automático aplicado al sector educativo permite la mejora de la eficiencia operativa de las instituciones educativas y también facilita un aprendizaje más efectivo para los estudiantes, contribuyendo en la mejora continua del sistema educativo en su conjunto.

Utilizando datos demográficos, académicos y de comportamiento, los modelos de Machine Learning pueden predecir qué estudiantes tienen mayor riesgo de abandonar sus estudios. Esto permite a las instituciones educativas intervenir de manera proactiva, ofreciendo apoyo personalizado y estrategias de retención para mejorar la tasa de graduación.

2.3. Estado del arte

En el ámbito de la gestión educativa, existen actualmente diversas herramientas que han sido desarrolladas para la explotación de información proveniente de sistemas de gestión del aprendizaje (LMS, por sus siglas en inglés, Learning Management System), cada una ofreciendo capacidades para mejorar la experiencia educativa y optimizar los procesos administrativos.

IntelliBoard: Esta plataforma de analítica del aprendizaje permite a educadores y administradores visualizar datos y realizar análisis detallados en tiempo real a través de paneles de control y reportes. Utiliza técnicas de aprendizaje automático para predecir resultados académicos, proporcionando información que puede ser utilizada para ajustar estrategias educativas y mejorar el rendimiento estudiantil. Es conocida por su interfaz intuitiva y su capacidad para integrarse con múltiples sistemas [5].

Learning Pool: Es una herramienta especializada en soluciones de aprendizaje digital y analítica. Permite a las instituciones recopilar, analizar y evaluar datos de manera efectiva. Sus herramientas son útiles para evaluar la efectividad de los programas educativos [8].

Watershed: Se centra en la captura y análisis detallado de datos de aprendizaje. Esta plataforma es valorada por su capacidad para ofrecer una comprensión profunda del progreso estudiantil a lo largo del tiempo y la identificación de áreas de mejora mediante análisis comparativos [29].

Blackboard: Es una plataforma LMS que ha evolucionado para incluir funcionalidades avanzadas de analítica. Se ha centrado en herramientas predictivas que utilizan datos históricos para prever resultados académicos y mejorar la retención estudiantil [2].

Open Education Analytics (OEA): Microsoft ofrece esta iniciativa de código abierto. Es una plataforma gratuita para que las instituciones educativas utilicen los datos disponibles con el objetivo de mejorar la gestión y los resultados educativos. Facilita la integración de datos de diversas fuentes y proporciona herramientas para el análisis. Sin embargo, la licencia gratuita de OEA presenta algunas limitaciones, como el análisis predictivo o la integración con otros sistemas. Además, tiene un límite en la cantidad de datos que se pueden procesar y almacenar, y está fuertemente integrado con los servicios de Microsoft Azure, lo que puede generar costos significativos asociados con el uso de estos servicios en la nube. Esta dependencia también limita la flexibilidad de despliegue en otras plataformas [14].

A pesar de los avances, la implementación efectiva de Business Intelligence y Machine Learning en el ámbito educativo enfrenta desafíos significativos. La integración, la calidad de los datos y la capacidad técnica para implementar estas tecnologías de manera efectiva son algunos de los obstáculos presentes. Otro aspecto a tomar en cuenta es la accesibilidad, las herramientas disponibles mencionadas anteriormente conllevan un coste asociado o disponen de licencias gratuitas con funcionalidades limitadas. Esta es la razón principal por la que estas herramientas no se adaptan al objetivo planteado, ya que limitan la accesibilidad para muchas instituciones educativas. Este es el factor determinante para crear una herramienta de libre uso, flexible y personalizable.

Capítulo 3

Tecnologías

3.1. Tecnologías utilizadas

La elección de la herramienta de BI para desarrollar el proyecto fue un aspecto crucial, guiado por un análisis comparativo de las opciones disponibles en el mercado. El objetivo era seleccionar herramientas que cumplieran con varios requisitos clave.

Inicialmente, se realizó una investigación para identificar las herramientas más populares para este tipo de desarrollo. Los resultados de la investigación fueron plasmados en una tabla comparativa (Tabla 3.1) que destacaba las principales opciones:

Herramienta	Tipo	Integración con scripts	Costo	Licencia	Curva de aprendizaje	Escalabilidad
Power BI	BI suite	Python y R	Pago / Gratuito (con limitaciones)	Propietaria	Media	Alta
Tableau	BI suite	Python y R	Pago / Gratuito (con limitaciones)	Propietaria	Alta	Alta
Qlik	BI suite	Qlik Sense script	Pago / Gratuito (con limitaciones)	Propietaria	Media	Alta
Apache Superset	BI suite	Python	Open source	Apache 2.0	Alta	Alta
Shiny	Framework	R y Python	Open source	GPL	Alta	Media
Streamlit	Framework	Python	Open source	Apache 2.0	Baja	Media
Dash	Framework	Python, Julia, R y F#	Open source	MIT	Media	Media

Tabla 3.1: Tabla comparativa de herramientas BI.

Se evaluaron diversos factores, comenzando por el coste y las licencias, priorizando herramientas que permitieran el libre uso y distribución de la aplicación. Aquellas que requerían un costo significativo fueron descartadas por las limitaciones que presentaban.

Además del aspecto económico, la integración con scripts era esencial debido a la necesidad de incorporar un modelo predictivo. Se buscaba también asegurar la escalabilidad, un requisito fundamental dada la cantidad de datos que se manejarán en un entorno de producción.

La curva de aprendizaje fue otro factor determinante, considerando el tiempo limitado para familiarizarse con las herramientas seleccionadas. En este contexto, también se valoró la categorización de las herramientas en BI suites y frameworks. Mientras las BI suites disponen de una menor personalización y flexibilidad, los frameworks programáticos ofrecen una mayor capacidad de adaptación a las necesidades específicas.

Finalmente, tras un análisis detallado, Dash y Streamlit fueron las opciones que mejor se adaptaban. Para decidir cuál utilizar, se tomaron en cuenta consideraciones sobre soporte, comunidad, documentación y opciones de despliegue.

Aunque Streamlit mostraba un crecimiento rápido y una comunidad activa, sus servicios y soporte aún no estaban tan consolidados como los de Dash, debido a su mayor antigüedad y presencia en el mercado. La documentación de Streamlit no igualaba la profundidad encontrada en Dash. En términos de despliegue, si bien Streamlit ofrecía opciones viables, Dash proporcionaba una flexibilidad y alcance más amplios en términos de plataformas y servicios compatibles. Por estas razones, la decisión final recayó en Dash.

- **Dash Plotly:** Es un framework de código abierto desarrollado por Plotly, reconocido por su capacidad para construir aplicaciones web interactivas y analíticas en Python. Ideal para desarrollar visualizaciones de datos y paneles de control que pueden compartirse y utilizarse a través de navegadores web [19].
- **Scikit-Learn:** Biblioteca de aprendizaje automático de código abierto para Python. Ofrece una amplia gama de herramientas y algoritmos para tareas de aprendizaje supervisado y no supervisado. Proporciona implementaciones optimizadas de algoritmos de clasificación, regresión, clustering y más [23].
- **SQLAlchemy:** Es una biblioteca de Python diseñada para facilitar el trabajo con bases de datos relacionales. Permite establecer la conexión entre la base de datos y la aplicación [24].
- **Faker:** Biblioteca de código abierto para Python, permite crear datos falsos, especialmente útil para pruebas y desarrollo [22].
- **PostgreSQL:** Sistema gestor de bases de datos relacionales y objeto-relacional, de código abierto, una opción robusta y eficiente para la gestión de los datos [20].
- **PgAdmin:** Herramienta de administración de bases de datos, de código abierto, diseñada principalmente para PostgreSQL, aunque también es compatible con otras bases de datos relacionales. Proporciona una interfaz gráfica intuitiva y funcionalidades avanzadas [16].
- **Git:** Sistema de control de versiones distribuido de código abierto, diseñado para gestionar proyectos de software de manera eficiente y colaborativa [4].
- **Visual Studio Code:** Editor de código fuente, conocido por su ligereza, velocidad y extensibilidad. Soporta una amplia variedad de lenguajes de programación y proporciona características avanzadas [11].

3.2. Análisis de la herramienta BI

Para garantizar que la aplicación se implementara de manera efectiva en un entorno de producción, fue fundamental tener en cuenta una serie de elementos importantes. Estos elementos incluyen la facilidad de mantenimiento y soporte continuo, la escalabilidad para manejar grandes cantidades de datos y el despliegue de la aplicación.

3.2.1. Funcionamiento

Dash utiliza React.js para la interfaz de usuario y Flask como servidor web, proporcionando una solución robusta y flexible.

React.js es una biblioteca JavaScript que se utiliza para crear interfaces de usuario. La descripción de los componentes de la interfaz en Python se traduce automáticamente a los componentes de React en Dash. Esto facilita la creación de interfaces interactivas y dinámicas sin tener que escribir código en JavaScript.

Flask es un microframework web para Python que ofrece una base sólida y ligera para el backend de aplicaciones web. Controla las solicitudes HTTP, la gestión de sesiones y otras características del servidor en Dash. La lógica del backend, las rutas y callbacks necesarias para la interacción de la interfaz de usuario son definidas por el código Python de Dash. Esto permite una integración fluida entre la lógica de negocio y la presentación.

Una aplicación Dash se basa en tres pilares fundamentales:

1. **Componentes de Dash:** Estos incluyen controles interactivos como botones, menús desplegables, etc. Estos elementos facilitan la interacción de los usuarios con la aplicación.
2. **Gráficos de Plotly:** Plotly es una biblioteca de gráficos interactivos que permite mostrar datos de manera efectiva. Los gráficos brindan a la interfaz de Dash una experiencia de usuario dinámica.
3. **Callbacks:** Los callbacks son una característica importante de Dash. Permiten la interactividad al conectar componentes de Dash y gráficos de Plotly. Definen cómo los cambios en los componentes de la interfaz deben afectar a otros componentes, permitiendo una respuesta dinámica y en tiempo real a las interacciones del usuario.

3.2.2. Escalabilidad

La herramienta debe manejar un número creciente de usuarios y altos volúmenes de datos. Dash proporciona las herramientas y la flexibilidad necesaria para construir aplicaciones web escalables. Sin embargo, la escalabilidad efectiva también depende del diseño, la implementación y la administración de la infraestructura de la aplicación y del servidor donde se ejecute [17]. Algunos de estos aspectos son los siguientes:

- **Manejo eficiente de datos:** Para manejar grandes volúmenes de datos, Dash permite integrarse con bases de datos SQL y NoSQL. Optimizar las consultas a la base de datos y utilizar técnicas de almacenamiento en caché son prácticas recomendadas para mejorar el rendimiento y la escalabilidad.
- **Escalabilidad horizontal y vertical:**
 - **Horizontal:** Puede escalar horizontalmente agregando instancias adicionales de la aplicación. Esto se puede lograr utilizando servicios en la nube con capacidades de autoescalado, como AWS Elastic Beanstalk o Kubernetes.
 - **Vertical:** También es posible escalar verticalmente añadiendo más recursos (CPU, memoria) a cada instancia de la aplicación Dash. Esto se traduce en una mejora en la capacidad de procesamiento y rendimiento de la aplicación.
- **Optimización de rendimiento:** Dash ofrece herramientas para optimizar el rendimiento, como la caché de datos y gráficos. Almacenar en caché resultados frecuentemente consultados puede reducir la carga en el servidor y mejorar los tiempos de respuesta, beneficiando la escalabilidad.

Las aplicaciones Dash están diseñadas para manejar múltiples usuarios concurrentes de manera eficiente. Para garantizar una experiencia de usuario óptima en un entorno multiusuario, se deben considerar varios aspectos:

- **Autenticación y autorización:** Implementar mecanismos de autenticación (por ejemplo, utilizando Flask-Login) para gestionar el acceso de los usuarios y asegurar que solo los usuarios autorizados puedan interactuar con la aplicación.
- **Manejo de sesiones:** Utilizar sesiones para mantener el estado del usuario a través de sus interacciones con la aplicación. Dado que Dash no maneja el estado de la sesión por defecto, es necesario implementar soluciones adicionales como el almacenamiento del estado en Redis o en una base de datos.

3.2.3. Despliegue

Para el despliegue de aplicaciones Dash existen varias opciones disponibles [18] :

- **Heroku:** Es una plataforma como servicio (PaaS) que simplifica el desarrollo, despliegue y gestión de aplicaciones en la nube. Destaca por su facilidad de uso y rápida configuración.
- **Docker:** Docker permite empaquetar la aplicación en contenedores, asegurando que se ejecute de manera consistente en diferentes entornos. Utilizando Docker Compose, se pueden definir y ejecutar aplicaciones multi-contenedor, facilitando la gestión de dependencias y la escalabilidad.
- **Servicios en la nube:** Plataformas como AWS Elastic Beanstalk y Google App Engine permiten desplegar aplicaciones Dash con configuraciones mínimas, proporcionando escalabilidad automática y gestión de recursos.
- **Servidor local:** Para despliegues en servidores locales, es posible utilizar servidores web como Nginx o Apache para gestionar el tráfico y redirigir las solicitudes a una aplicación Dash que se ejecute con un servidor WSGI como Gunicorn. Esta configuración proporciona flexibilidad y control total sobre el entorno de ejecución.
- **Dash Enterprise:** Es una plataforma de despliegue y gestión de aplicaciones Dash proporcionada por Plotly. Ofrece autenticación integrada, administración de usuarios, integración con CI/CD y opciones avanzadas de despliegue. Sin embargo, este es un servicio completamente de pago.

Capítulo 4

Desarrollo

En este capítulo se detalla el proceso de desarrollo de la aplicación, abarcando desde el análisis y estudio de la base de datos, la generación de datos sintéticos, el diseño de la visualización de los datos y la arquitectura de software, hasta la implementación de un modelo de aprendizaje.

4.1. Generación de datos

Los datos son indispensables para evaluar los resultados de cualquier herramienta o estudio. En este apartado, se describe detalladamente el proceso de análisis de los datos, desde el punto de partida hasta su generación final. Este proceso es fundamental para asegurar la calidad y la fiabilidad de los resultados obtenidos.

4.1.1. Diseño de base de datos

Antes de iniciar el presente proyecto, se llevó a cabo un estudio por parte del tutor académico sobre el diseño de la base de datos del proyecto *“Análisis de los factores explicativos del abandono de los estudios universitarios y acciones estratégicas para su mejora y prevención”*. Este análisis inicial incluyó la evaluación de la estructura existente. Partiendo de esta base sólida, el proyecto continuó con la línea de trabajo (Figura 4.1).

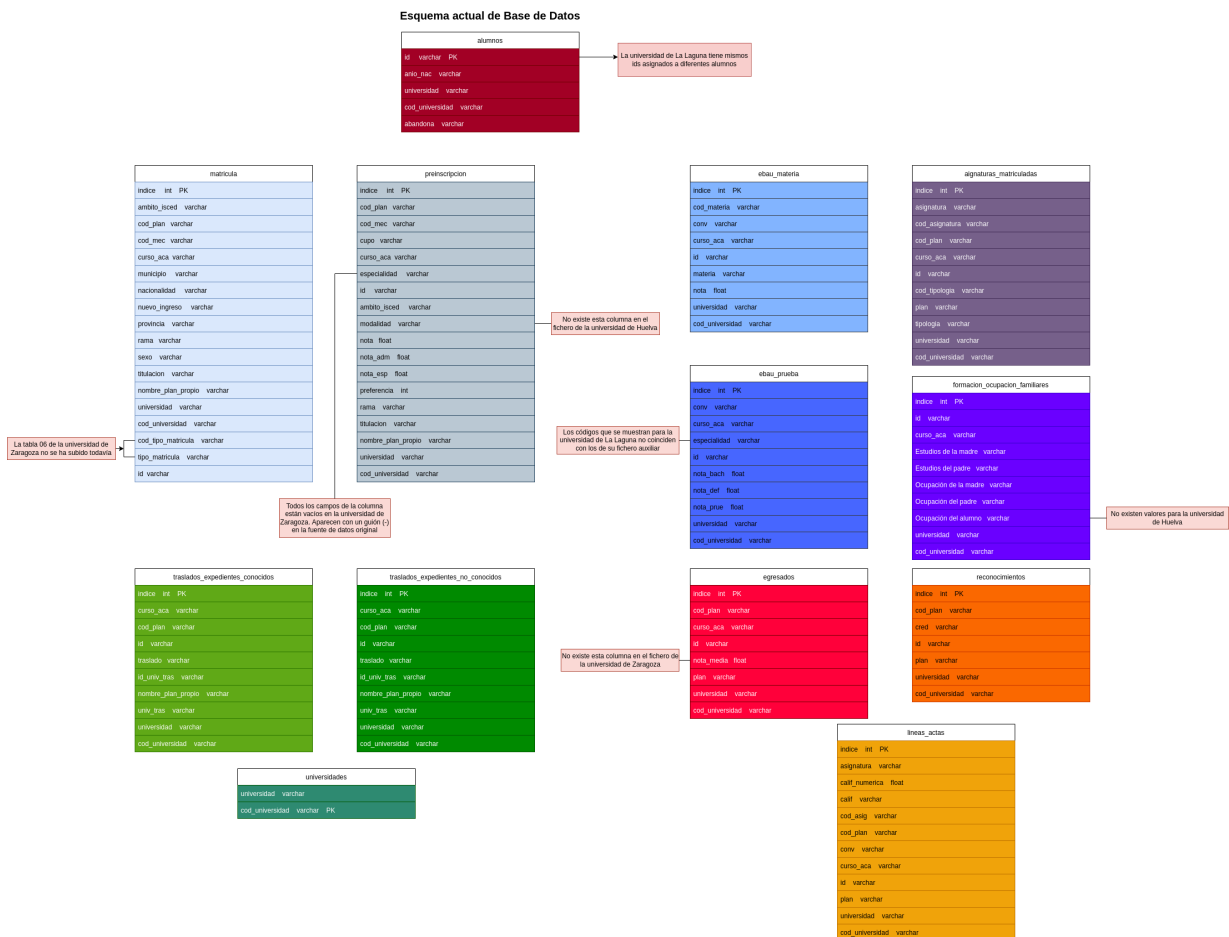


Figura 4.1: Esquema de la base de datos inicial (Origen proyecto ABACU).

Para el desarrollo de la aplicación se utilizaron la mayoría de las tablas del esquema proporcionado. Las tablas utilizadas fueron las siguientes:

- Alumnos:** Esta tabla representa la información básica relativa sobre los alumnos, como el año de nacimiento, la universidad a la que pertenecen y si abandonan la titulación. Es importante destacar que el campo abandona es un valor previamente calculado. Se considera que un alumno se encuentra en un estado de abandono cuando, tras matricularse por primera vez en un grado, no se ha vuelto a matricular durante dos cursos consecutivos ni ha titulado en un máximo de cuatro cursos desde el primero, acumulando dos años sin matrícula [13].
- Matrícula:** Representa la matrícula que debe tener cada alumno por período académico e incluye más detalles relacionados con el alumno y la titulación.
- Asignaturas matriculadas:** Contiene información sobre las asignaturas que cada alumno ha matriculado durante cada curso académico.

- **Líneas actas:** Dispone de los resultados académicos de los alumnos, incluyendo las asignaturas presentadas, las calificaciones obtenidas en cada convocatoria y otros detalles relevantes.
- **Ebau prueba:** Contiene las calificaciones obtenidas por cada alumno, referentes a la nota de acceso a la universidad.
- **Egresados:** Representa los alumnos egresados por universidad y la nota media que obtuvieron.

Durante el desarrollo de la aplicación se produjo la necesidad de incluir dos tablas adicionales referentes a dos perfiles diferentes (Figura 4.2):

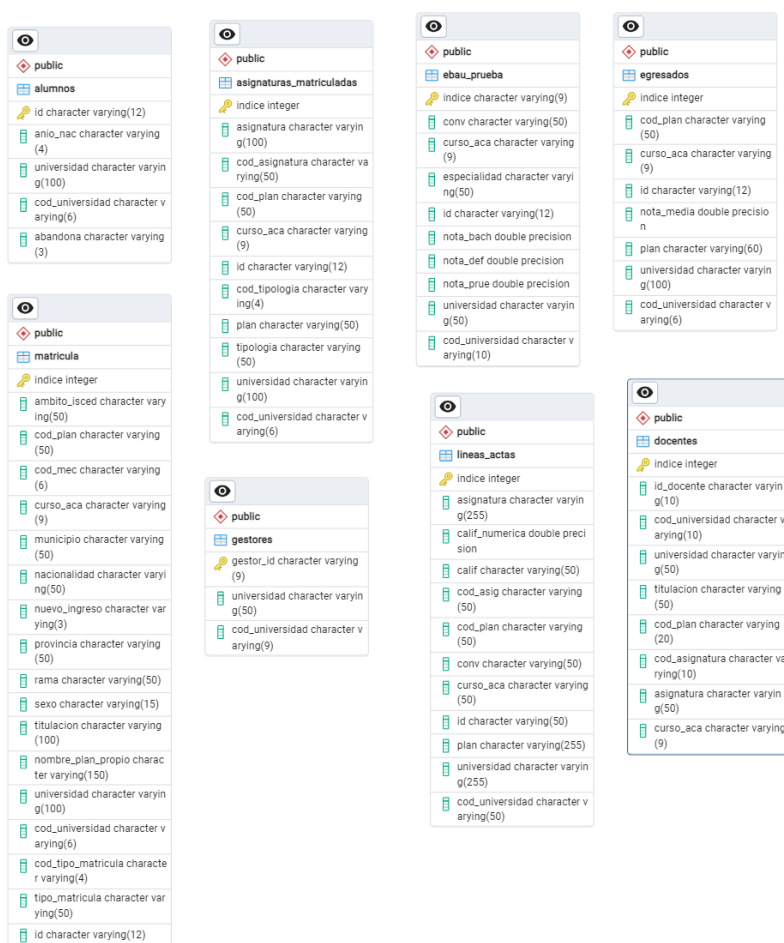


Figura 4.2: Esquema de base de datos utilizado.

- **Docentes:** Contiene información de los profesores y de cada una de las asignaturas que imparten en cada titulación.
- **Gestores:** Representa a los gestores por universidad. Los gestores son aquellos que tienen una responsabilidad de gestión, a nivel de titulación, escuelas y facultades, claustro y equipo rectoral, consejo social, etc.

4.1.2. Datos sintéticos

Debido a la imposibilidad de acceso sobre los datos reales del entorno académico, por motivos de protección de datos, se implementó una solución alternativa: desarrollar una herramienta de generación de datos sintéticos, con el objetivo de simular una versión realista de los datos originales.

La herramienta permite especificar la cantidad de alumnos, docentes y gestores que se quieren generar. A partir de la cantidad de alumnos generados, se determina el volumen de datos en las demás tablas. Posteriormente, se crean los conjuntos de datos para cada tabla, comenzando por los datos para la tabla "alumnos", seguido de "matricula", "asignaturas matriculadas", "lineas actas", "docentes", "ebau prueba", "egresados" y "gestores". Una vez generados los conjuntos de datos correspondientes, se exporta su contenido a los archivos CSV respectivos. Luego estos ficheros generados son almacenados en el directorio `src/data/csv`, listos para ser importados posteriormente a la base de datos.

Fue fundamental el uso de dos módulos para conseguir el objetivo, `faker` y `random`. `Faker` facilitó la creación de valores aleatorios para variables categóricas como identificadores, palabras aleatorias, nombres de ciudades y países. El módulo `random` permite obtener datos aleatorios de variables numéricas como las calificaciones de los alumnos, rangos entre valores y la selección aleatoria de valores sobre los diccionarios creados con información relativa a las universidades, ramas del conocimiento, titulaciones y asignaturas.

Durante la fase de desarrollo, el uso de los datos generados fue esencial para realizar todas las pruebas correspondientes y permitió visualizar los resultados de la herramienta. Estos datos sintéticos crearon un entorno controlado para evaluar la aplicación, asegurando que todos los componentes y las consultas a la base de datos funcionaran correctamente.

4.2. Diseño de visualización de datos

Realizar un estudio del diseño de visualización de los datos ayuda a evitar errores que pudieran perjudicar el proceso de desarrollo. Por esta razón, se decidió diseñar mockups para tener un primer acercamiento y formalizar una idea visual del proyecto.

En esta primera versión, se plantearon y abordaron las siguientes cuestiones: ¿A qué tipo de usuario iba destinada la herramienta? ¿Qué tipo de información es conveniente representar según el usuario?

En primer lugar, como respuesta a la primera pregunta, se llegó a la conclusión de que los usuarios interesados eran los siguientes:

- **Alumnos:** Considerados como los principales beneficiarios de la herramienta, ya que son el foco central de las investigaciones. La herramienta está diseñada para proporcionarles información relevante que pueda mejorar su experiencia educativa.
- **Docentes:** Los docentes juegan un papel importante. Están directamente involucrados con los estudiantes y pueden utilizar la herramienta para obtener una perspectiva sobre el rendimiento de sus alumnos, identificar áreas de mejora y ajustar sus métodos de enseñanza.
- **Gestores educativos:** Estos usuarios necesitan una visión global y detallada del rendimiento educativo a nivel institucional. Esto les permitirá tomar decisiones informadas, implementar estrategias educativas eficaces y supervisar el progreso hacia los objetivos institucionales.

La respuesta a la segunda pregunta se explica con detalle en los siguientes subapartados:

4.2.1. Perfil alumnado

En un primer análisis, se tomó en cuenta la importancia de que cada alumno pueda visualizar sus propios resultados a través de un resumen y gráficos interactivos (Figura 4.3). La información a representar fue la siguiente:



Figura 4.3: Primera versión mockup. Perfil alumnado.

- **Evolución de la nota media:** Permite a los estudiantes visualizar la evolución de su rendimiento académico a lo largo del tiempo. Al observar las tendencias en sus calificaciones, los alumnos pueden identificar períodos de buenos o malos resultados, lo que les permite tomar acciones correctivas o mantener las estrategias de estudio.
- **Número de asignaturas aprobadas:** Mostrar el número de asignaturas aprobadas brinda a los estudiantes una visión clara del progreso hacia la finalización de los estudios. Permite mantener la motivación y para la planificación de futuros cuatrimestres.
- **Nota por asignaturas:** Los estudiantes pueden utilizar esta información para enfocar sus esfuerzos de estudio en las asignaturas más complejas y para buscar ayuda.

Además, se querían mostrar representaciones comparativas con los resultados colectivos de otros estudiantes, en este caso interactuando con filtros como curso académico y asignaturas.

- **Evolución de la nota media general:** Comparar la evolución de la nota media general con la de otros estudiantes proporciona un contexto adicional sobre el rendimiento académico del alumno.

- **Nota media por asignatura y curso:** Esto puede destacar si una asignatura específica está resultando más difícil para todos los estudiantes, o si el alumno tiene un rendimiento diferente al promedio.
- **Porcentaje de aprobados por curso académico:** El porcentaje de aprobados por curso académico proporciona una visión general de la tasa de éxito en diferentes cursos.

4.2.2. Perfil docente

En el perfil docente era conveniente mostrar los resultados del rendimiento académico de las asignaturas impartidas y visualizar la evolución temporal de los resultados. A continuación, se proporciona una visión para cada uno de los gráficos y datos representados (Figura 4.4).

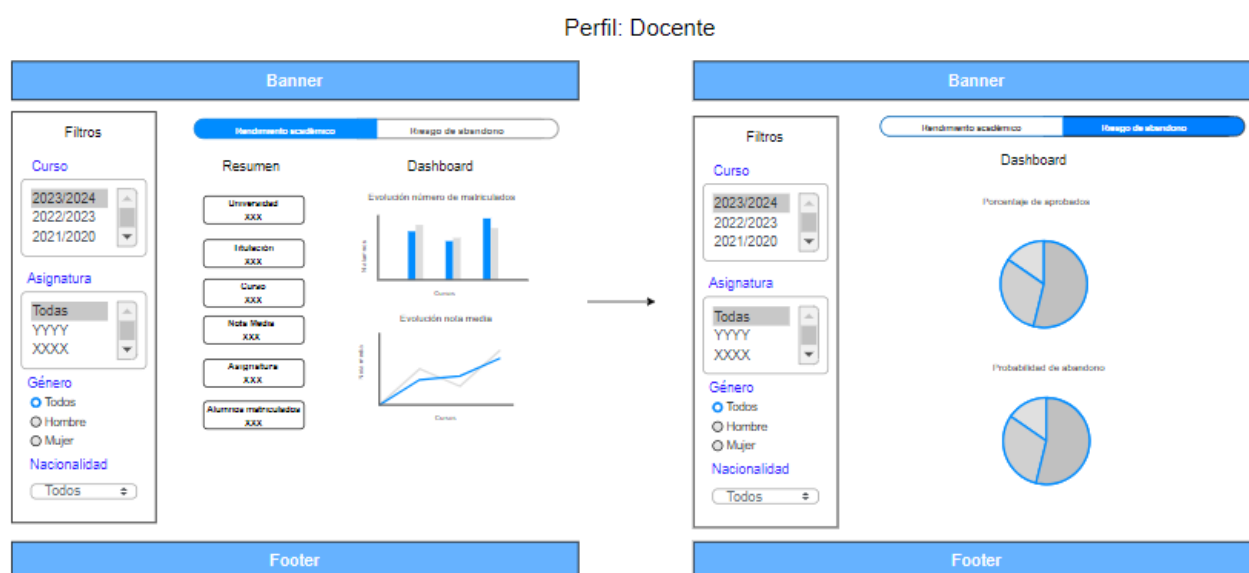


Figura 4.4: Primera versión mockup. Perfil docente.

Las representaciones destacadas fueron las siguientes:

- **Evolución del número de matriculados:** Es una métrica importante porque puede indicar la percepción de dificultad de la asignatura.
- **Evolución nota media:** Permite a los docentes observar cómo ha cambiado el rendimiento promedio de los estudiantes en sus asignaturas a lo largo de diferentes periodos.

Además, a través de otra pestaña, se buscaba representar los resultados relacionados al riesgo de abandono académico:

- **Porcentaje de aprobados:** Los docentes pueden usar esta información para evaluar la tasa de éxito en sus cursos e identificar áreas que necesiten refuerzo adicional.
- **Probabilidad de abandono:** Se puede utilizar esta información para identificar cursos con altas tasas de abandono e investigar las causas.

4.2.3. Perfil gestor

En el perfil gestor, en lugar de representar los resultados en base a las asignaturas, en este caso se quería representar los resultados por titulaciones, teniendo una representación más amplia de los resultados a nivel institucional (Figura 4.5).

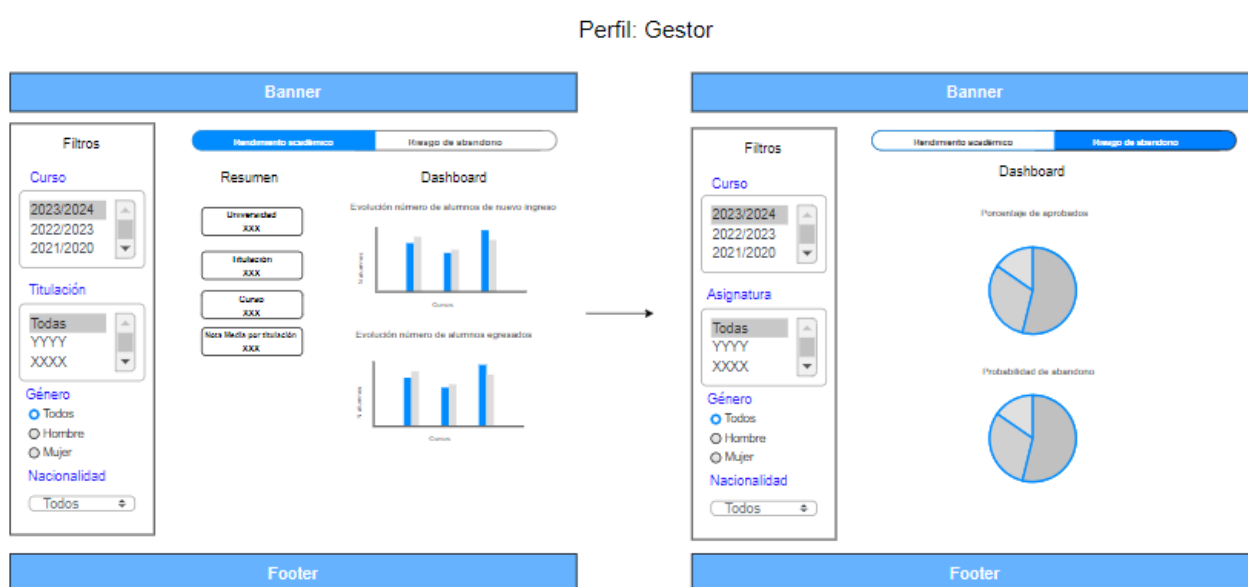


Figura 4.5: Primera versión mockup. Perfil gestor.

En el apartado “Rendimiento académico” se muestra la siguiente información:

- **Evolución del número de alumnos de nuevo ingreso:** Este gráfico muestra cómo ha cambiado la cantidad de nuevos estudiantes matriculados en las titulaciones a lo largo del tiempo. Es un indicador útil para evaluar la atractividad de los programas académicos y planificar estrategias de captación.
- **Evolución del número de alumnos egresados:** Muestra el número de estudiantes que han completado sus estudios y se han graduado en cada titulación. Este indicador permite evaluar el éxito y la eficacia de los programas educativos.

En el apartado de “Riesgo de abandono” se buscaba mostrar las siguientes representaciones:

- **Porcentaje de aprobados:** Proporciona una visión del porcentaje de estudiantes que aprobaron las titulaciones. Es información clave para medir la calidad y el rendimiento académico a nivel institucional.
- **Probabilidad de abandono:** Los gestores pueden usar esta información para identificar titulaciones con altas tasas de abandono, analizar los factores contribuyentes y diseñar intervenciones específicas para reducir el abandono.

4.2.4. Validación del diseño

Se realizó una validación con expertos (proyecto ABACU), en la que se presentaban los mockups realizados, se recogían su valoración y propuestas de mejora.

Se continuó este procedimiento iterativamente hasta alcanzar el diseño final de los paneles de visualización, mejorando significativamente la calidad. En el capítulo 5 del documento se mostrarán los resultados finales obtenidos (5).

4.3. Diseño y arquitectura de software

El proyecto está basado en una arquitectura de componentes (Figura 4.6), un enfoque de diseño de software que se centra en la creación de sistemas mediante la composición de unidades de software más pequeñas y autónomas, llamadas componentes. Cada componente representa una parte funcional específica de la aplicación y puede ser desarrollado, probado y mantenido de manera independiente. Este enfoque ofrece varias ventajas:

- **Reutilización de código:** Los componentes pueden ser reutilizados en diferentes partes de la aplicación o en proyectos futuros, lo que reduce el tiempo y el esfuerzo de desarrollo.
- **Facilidad de mantenimiento:** La modularidad permite realizar cambios o actualizaciones en un componente sin afectar otros componentes, facilitando la gestión y el mantenimiento del software.
- **Escalabilidad:** La aplicación puede expandirse fácilmente añadiendo nuevos componentes o modificando los existentes sin necesidad de rediseñar toda la arquitectura.

La estructura del proyecto se divide de la siguiente forma:

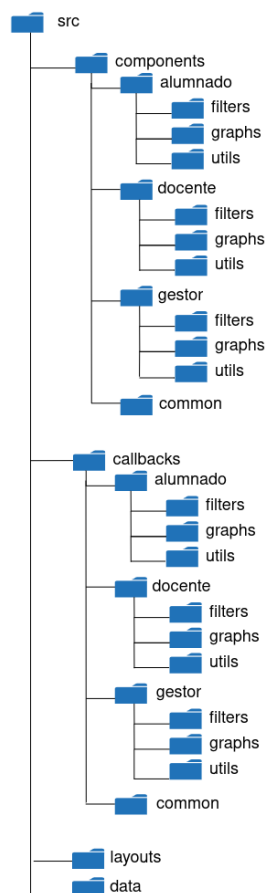


Figura 4.6: Estructura de directorios del proyecto.

- **src/components:** Representa los componentes de la aplicación para cada uno de los roles presentes: alumnado, docente, gestor. A su vez están agrupados por filtros, gráficos y otros tipos de componentes como selectores, pestañas, etc.
- **src/callbacks:** Contiene las funciones encargadas de gestionar cada uno de los eventos que disparan los componentes. Representa la lógica y el tratamiento de los datos para representar las visualizaciones e interacciones.
- **src/layout:** Contiene las vistas disponibles para cada uno de los roles.
- **src/data:** En este directorio se encuentran todos los recursos relacionados con los datos, configuraciones, conexión con la base de datos, consultas y generación de datos.

4.4. Modelo de aprendizaje automático

Implementar un modelo de aprendizaje automático permite identificar tempranamente a los estudiantes que están en riesgo de abandonar sus estudios. Esta predicción facilita la intervención proactiva, lo que significa que se pueden implementar estrategias de apoyo personalizadas y específicas para cada estudiante, con el propósito de poder aumentar las tasas de retención y éxito académico.

Por esta razón, se consideró oportuna la implementación de un modelo de aprendizaje automático en la herramienta para determinar la probabilidad de abandono de cada estudiante que se encuentra en un estado activo, es decir, que no ha abandonado la titulación ni ha finalizado los estudios.

Para implementar un modelo de aprendizaje automático se realizan varios pasos sistemáticos:

- Definición del problema.
- Recogida de datos.
- Procesamiento de los datos.
- Selección del algoritmo.
- Entrenamiento del modelo.
- Evaluación del modelo.

4.4.1. Recogida de datos

Los datos utilizados en este modelo incluyen información demográfica y académica. Estos datos son recopilados y preprocesados para garantizar su integridad y utilidad en el modelo de predicción. La carga y preparación inicial de los datos se realiza mediante una función que obtiene información de la base de datos.

4.4.2. Procesamiento de los datos

- **Creación de nuevas variables:** Se calcula la edad actual del estudiante utilizando su año de nacimiento y el año actual.
- **Codificación de variables:** La variable objetivo (abandona) se codifica en formato binario, donde 1 indica que el estudiante abandonó y 0 indica que no lo hizo.
- **Manejo de valores faltantes:** Se rellenan los valores faltantes en las columnas numéricas con la media y en las columnas categóricas con un valor constante.

- **Interacciones polinómicas:** Se crean características polinómicas para capturar interacciones entre las características existentes, lo que puede mejorar el rendimiento del modelo.
- **Selección de características:** Se seleccionaron las variables de la base de datos que podrían aportar más valor al modelo y mejorar la precisión de los resultados. Se utilizaron datos demográficos y resultados académicos: “nacionalidad”, “sexo”, “titulación”, “nota_def_acceso”, “nota_media” y “anio_nac”.
 - **Nacionalidad:** Las diferencias culturales, lingüísticas y los sistemas educativos previos pueden afectar la probabilidad de que los estudiantes abandonen.
 - **Sexo:** Estudios han mostrado que existen diferencias en las tasas de abandono entre hombres y mujeres, lo que hace que esta variable sea relevante para el modelo [13].
 - **Titulación:** El tipo de titulación que el alumno está cursando puede influir en la tasa de abandono, ya que algunas titulaciones pueden ser más exigentes.
 - **Nota de acceso:** La nota de acceso a la universidad es un indicador del rendimiento académico previo del alumno. Esto puede predecir la probabilidad de éxito o abandono en la universidad.
 - **Nota media:** La nota media actual del alumno es un indicador de rendimiento académico continuo que puede influir en el abandono académico.
 - **Año de nacimiento:** La edad puede influir en la probabilidad de que un estudiante abandone la escuela. Los alumnos con más edad pueden tener más responsabilidades y los más jóvenes menos experiencia para adaptarse.
- **Escalado y codificación:** Las características numéricas se escalan para tener una media de 0 y una desviación estándar de 1, mientras que las categóricas se codifican utilizando codificación one-hot para convertirlas en variables numéricas.
- **División de datos:** Los datos se dividen en conjuntos de entrenamiento y prueba, asegurando que el modelo pueda ser evaluado adecuadamente.
- **Balanceo de clases:** Se aplica la técnica de sobremuestreo SMOTE para balancear las clases en el conjunto de entrenamiento, dado que el abandono puede ser una clase minoritaria (Código 4.1).

```

1
2 current_year = datetime.now().year
3 data['edad_actual'] = current_year - data['anio_nac']
4
5 # Codificar variable objetivo
6 data['abandona'] = data['abandona'].map({'si': 1, 'no': 0})
7
8
9 columnas_categoricas = ['nacionalidad', 'sexo', 'titulacion']
10 columnas_numericas = ['nota_def_acceso', 'nota_media', 'edad_actual']
11
12 # Preprocesamiento
13 preprocessor = ColumnTransformer(
14     transformers=[
15         ('num', Pipeline([('imputer', SimpleImputer(strategy='mean')), ('scaler',
16             ↪ StandardScaler())])), columnas_numericas),
17         ('cat', Pipeline([('imputer', SimpleImputer(strategy='constant',
18             ↪ fill_value='Desconocido')),
19             ('encoder', OneHotEncoder(handle_unknown='ignore'))])),
20             ↪ columnas_categoricas)
21     ])
22
23 # Dividir datos en características y variable objetivo
24 X = data[columnas_categoricas + columnas_numericas]
25 y = data['abandona']
26 ids = data['id']
27
28 # Dividir los datos en conjuntos de entrenamiento y prueba
29 X_train, X_test, y_train, y_test, id_train, id_test = train_test_split(X, y, ids,
30     ↪ test_size=0.2, random_state=42, stratify=y)
31
32 X_train = preprocessor.fit_transform(X_train)
33 X_test = preprocessor.transform(X_test)
34
35 smote = SMOTE(random_state=42)
36 X_train, y_train = smote.fit_resample(X_train, y_train)
37
38 # Crear interacciones polinómicas
39 poly = PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)
40 X_train = poly.fit_transform(X_train)
41 X_test = poly.transform(X_test)
42

```

Listing 4.1: Código del procesamiento de datos.

4.4.3. Selección del algoritmo

Los algoritmos utilizados se enmarcan dentro del aprendizaje supervisado. El modelo se entrena utilizando un conjunto de datos etiquetados. En este caso, las etiquetas indican si un estudiante ha abandonado o no. El aprendizaje supervisado es adecuado para este problema porque permite al modelo aprender patrones a partir de datos históricos para hacer predicciones precisas sobre futuros casos de abandono (Código 4.2). Se utilizan varios algoritmos de clasificación:

Se han seleccionado varios algoritmos de clasificación por su capacidad para manejar diferentes aspectos y complejidades del problema:

- **Random Forest:** Este algoritmo utiliza múltiples árboles de decisión para mejorar la precisión y reducir el riesgo de sobreajuste. Cada árbol se entrena con un subconjunto diferente de los datos, lo que hace robusto para las relaciones no lineales.
- **Logistic Regression:** Un modelo lineal que es eficiente y simple, ideal para problemas de clasificación binaria como este.
- **KNeighbors:** Este método tiene la capacidad de capturar relaciones locales en los datos, lo que lo hace especialmente valioso en situaciones donde hay patrones complejos y no lineales.
- **AdaBoost:** Un algoritmo de boosting que combina múltiples clasificadores débiles, dándoles más peso a aquellos que corrigen errores de predicciones anteriores.
- **XGBoost:** Una versión optimizada de Gradient Boosting, conocida por su rendimiento y escalabilidad en una variedad de problemas de clasificación.
- **LightGBM:** Destaca por su velocidad y eficiencia, especialmente al manejar grandes conjuntos de datos. Su capacidad para trabajar eficientemente con datos desequilibrados lo hace muy versátil.

Cada uno de estos algoritmos tienen características únicas que se ajustan a diferentes aspectos del problema, desde manejar datos desequilibrados hasta capturar relaciones complejas entre variables.

```

1
2 # Pipelines para cada modelo
3 modelos = {
4     'RandomForest': RandomForestClassifier(random_state=42, n_jobs=-1),
5     'LogisticRegression': LogisticRegression(max_iter=1000, random_state=42),
6     'KNeighbors': KNeighborsClassifier(n_jobs=-1),
7     'AdaBoost': AdaBoostClassifier(random_state=42),
8     'XGBoost': xgb.XGBClassifier(random_state=42, use_label_encoder=False,
9     ↪ eval_metric='logloss'),
10    'LightGBM': lgb.LGBMClassifier(random_state=42, verbose=-1,
11    ↪ force_row_wise=True)
12 }
13
14 # Hiperparámetros para búsqueda
15 param_grids = {
16     'RandomForest': {
17         'n_estimators': [100, 200],
18         'max_depth': [10, 20],
19         'min_samples_split': [2, 5],
20         'min_samples_leaf': [1, 2]
21     },
22     'LogisticRegression': {
23         'C': [0.01, 0.1, 1.0],
24         'solver': ['liblinear', 'lbfgs']
25     },
26     'KNeighbors': {
27         'n_neighbors': [3, 5, 7],
28         'weights': ['uniform', 'distance']
29     },
30     'AdaBoost': {
31         'n_estimators': [50, 100],
32         'learning_rate': [0.01, 0.1]
33     },
34     'XGBoost': {
35         'n_estimators': [100, 200],
36         'learning_rate': [0.01, 0.1],
37         'max_depth': [3, 5]
38     },
39     'LightGBM': {
40         'n_estimators': [100, 200],
41         'learning_rate': [0.01, 0.1],
42         'max_depth': [3, 5]
43     }
44 }

```

Listing 4.2: Código de selección de algoritmos.

4.4.4. Entrenamiento y evaluación de los modelos

El uso de una variedad de algoritmos de clasificación permite:

- **Comparación de modelos:** Evaluar cuál algoritmo proporciona las mejores predicciones para el problema específico.
- **Ensemble learning:** Combinar múltiples modelos (como en el caso del Voting Classifier) para mejorar la precisión y robustez del modelo final.

Para cada uno de estos algoritmos, se realiza una búsqueda exhaustiva de hiperparámetros utilizando GridSearchCV. Esto permite identificar los mejores parámetros que optimizan el rendimiento del modelo.

La evaluación de los modelos se realiza utilizando varios indicadores de rendimiento en el conjunto de prueba, tales como:

- Accuracy.
- Precision.
- Recall.
- F1 Score.
- ROC AUC.
- Matriz de Confusión.

Posteriormente, se entrena un clasificador de votación que combina los mejores modelos de cada algoritmo para mejorar la precisión general. El uso de un clasificador de votación asegura que se aprovechen las fortalezas de múltiples algoritmos, proporcionando una predicción robusta (Código 4.3). Después de crear el ensemble, se hace una evaluación del modelo para comparar estos resultados con los valores obtenidos anteriormente. Los resultados de la evaluación de los modelos se analizan en el capítulo 5.2.

Además, se realiza una validación cruzada para obtener una estimación más robusta del rendimiento del modelo, asegurando que este sea generalizable. Finalmente, el modelo entrenado se guarda para su posterior uso.

```

1
2 # Buscar los mejores hiperparámetros y entrenar modelos
3 best_estimators = {}
4 metrics = {}
5
6 with tqdm(total=len(modelos), desc="Entrenando modelos") as pbar:
7     for key, model in modelos.items():
8         grid_search = GridSearchCV(model, param_grids[key],
9             ↪ cv=StratifiedKFold(n_splits=5), scoring='accuracy', n_jobs=-1)
10        grid_search.fit(X_train, y_train)
11        best_estimators[key] = grid_search.best_estimator_
12
13 # Evaluar modelo
14 y_pred = grid_search.predict(X_test)
15 metrics[key] = {
16     'accuracy': accuracy_score(y_test, y_pred),
17     'precision': precision_score(y_test, y_pred),
18     'recall': recall_score(y_test, y_pred),
19     'f1_score': f1_score(y_test, y_pred),
20     'roc_auc': roc_auc_score(y_test, y_pred)
21 }
22 print("")
23 print(f"Resultados para {key}:")
24 print(f"Accuracy: {metrics[key]['accuracy']:.2f}")
25 print(f"Precision: {metrics[key]['precision']:.2f}")
26 print(f"Recall: {metrics[key]['recall']:.2f}")
27 print(f"F1 Score: {metrics[key]['f1_score']:.2f}")
28 print(f"ROC AUC: {metrics[key]['roc_auc']:.2f}")
29 print("Matriz de Confusión:")
30 print(confusion_matrix(y_test, y_pred))
31 print("")
32
33 pbar.update(1)
34
35 # Selecciona los mejores modelos para el ensemble
36 mejores_modelos = sorted(metrics.items(), key=lambda item: item[1]['f1_score'],
37     ↪ reverse=True)[:3]
38 mejores_estimadores = [(key, best_estimators[key]) for key, _ in mejores_modelos]
39
40 # Crear VotingClassifier con los mejores modelos
41 voting_clf = VotingClassifier(
42     estimators=mejores_estimadores,
43     voting='soft',
44     n_jobs=-1
45 )

```

Listing 4.3: Código entrenamiento del modelo.

Capítulo 5

Resultados

En la sección de resultados se refleja la fase final del proyecto, incluyendo la visualización de los datos y los resultados obtenidos del modelo de aprendizaje automático. Para garantizar una mejor comprensión del funcionamiento de la herramienta, se utilizó Doxygen para generar la documentación del código (Repositorio del proyecto [15]).

5.1. Visualización y funcionalidades

En el siguiente apartado, se presentan los resultados obtenidos a partir de un diseño de visualización de los datos realizado previamente. Además, se explican los cambios finales implementados y cómo han contribuido al logro de los objetivos planteados inicialmente.

5.1.1. Perfil alumnado

Partiendo de la idea general, se plantearon cambios y funcionalidades adicionales que permiten mejorar la interacción con el usuario y aportar valor a los resultados obtenidos.

En el perfil alumnado se realizaron tres divisiones correspondientes a los tipos de representaciones, a través de pestañas. Además, de un panel informativo en la parte izquierda con el resumen de la información del usuario y los filtros que interactúan con los gráficos. Las representaciones se dividen de la siguiente forma:

- **Expediente académico personal:** En este apartado, el propio alumno puede consultar su información personal referente a sus resultados académicos. Se dispone de un resumen con información de interés: universidad, titulación, nombre del alumno, nota media actual, estado del expediente y el porcentaje de probabilidad de abandono. Además, se incluyen dos filtros que interactúan directamente con los gráficos del cuadro de mando: titulación y curso académico (Figura 5.1).

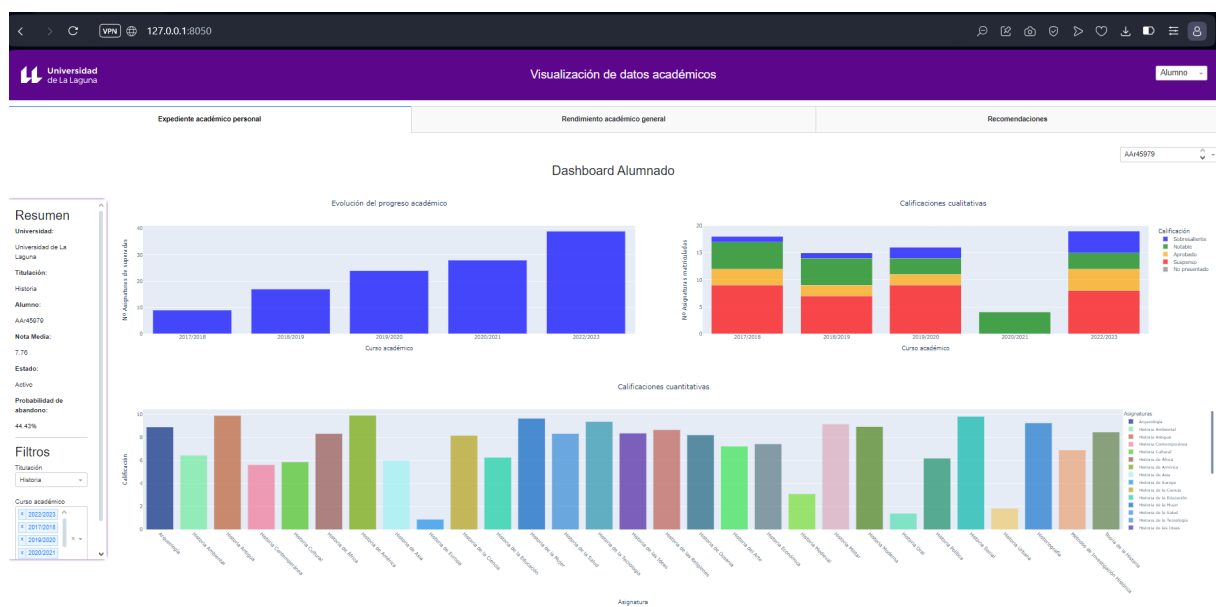


Figura 5.1: Perfil alumnado. Pestaña expediente académico personal.

Con respecto a los mockups diseñados en un principio, se realizaron cambios notables a nivel de diseño de interfaces y relacionados al tipo de representaciones que puede ser convenientes para los alumnos. Las representaciones gráficas que se visualizan son las siguientes:

- **Evolución del progreso académico:** A través de un gráfico de barras, se muestra el número de asignaturas superadas por curso académico de forma acumulativa. Esta representación es útil para que los estudiantes visualicen su progreso a lo largo del tiempo, identificando períodos de mayor y menor rendimiento académico (Figura 5.2).
- **Calificaciones cualitativas:** Permite a los estudiantes ver una distribución clara de sus calificaciones cualitativas, proporcionando una visión rápida de su rendimiento general en cada curso (Sobresaliente, Notable, Aprobado, Suspen­dido y No presentado). Se utiliza una leyenda que sirve para indicar el color que representa cada calificación y para filtrar las calificaciones gracias a la funcionalidad que dispone (Figura 5.3).

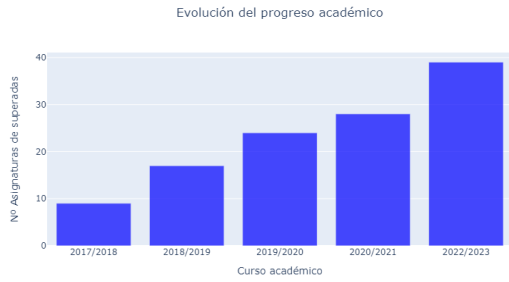


Figura 5.2: Gráfica. Evolución del progreso académico.

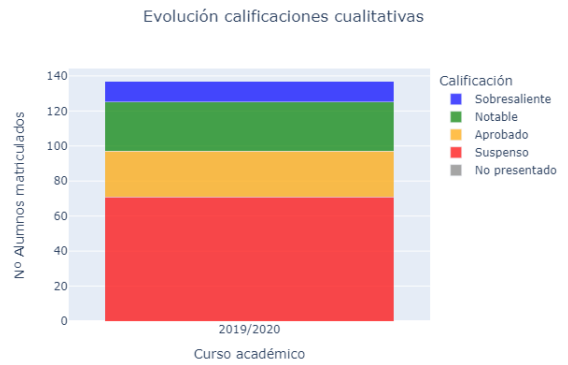


Figura 5.3: Gráfica. Evolución calificaciones cualitativas.

- **Calificaciones cuantitativas:** Para representar esta información se utiliza un gráfico de barras. Cada barra representa una asignatura matriculada y la calificación numérica obtenida por el alumno (Figura 5.4).

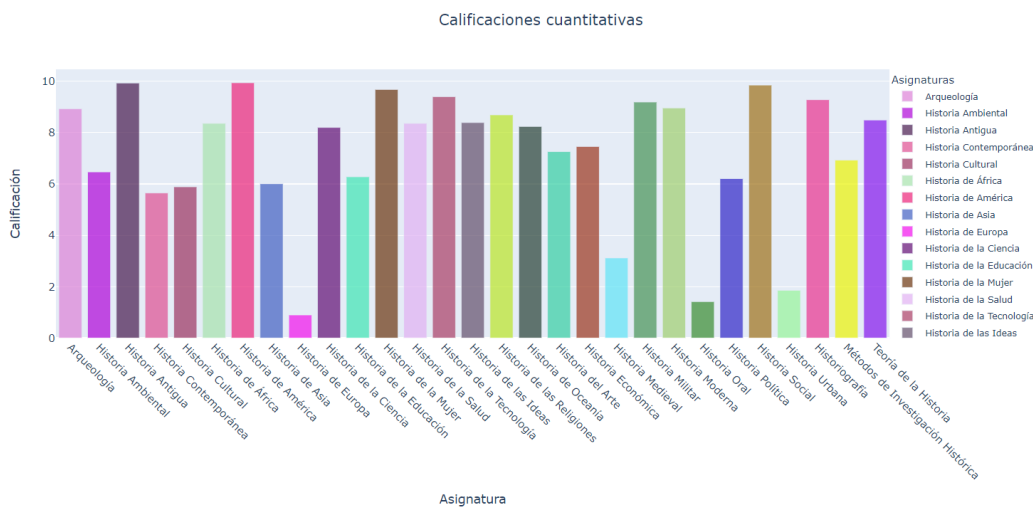


Figura 5.4: Gráfica. Calificaciones cuantitativas.

- **Tasa de rendimiento:** Está representado por un gráfico de barras en horizontal. Esta métrica es esencial para evaluar el rendimiento general del estudiante, con respecto a la superación de asignaturas matriculadas (Figura 5.5):

$$\text{Rendimiento académico} = \frac{\text{Número de asignaturas superadas}}{\text{Número de asignaturas matriculadas}} * 100$$

[12]

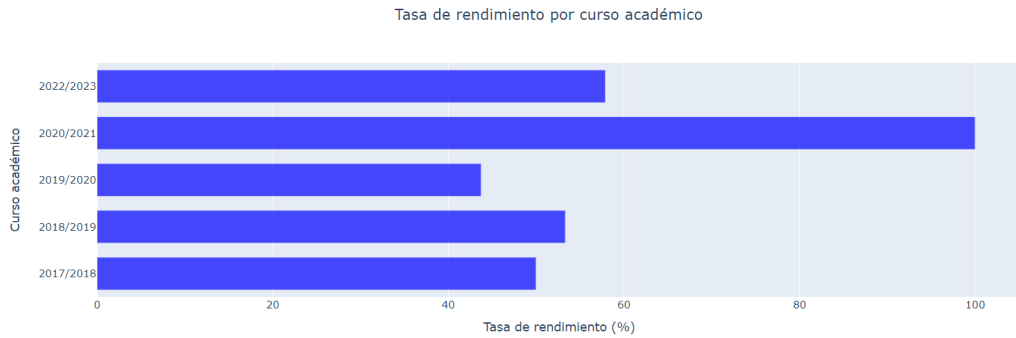


Figura 5.5: Gráfica. Tasa de rendimiento académico.

- Rendimiento académico general:** En esta sección se presenta información general del alumnado que cursa asignaturas en común con el alumno en cuestión. Este enfoque permite comparar los resultados del alumno con los de sus compañeros y permite identificar el estado de riesgo en que se encuentra (Figura 5.6).

En cuanto al panel informativo, se comparte el resumen de información de la pestaña mencionada anteriormente. Además, se añade el filtro de asignaturas.

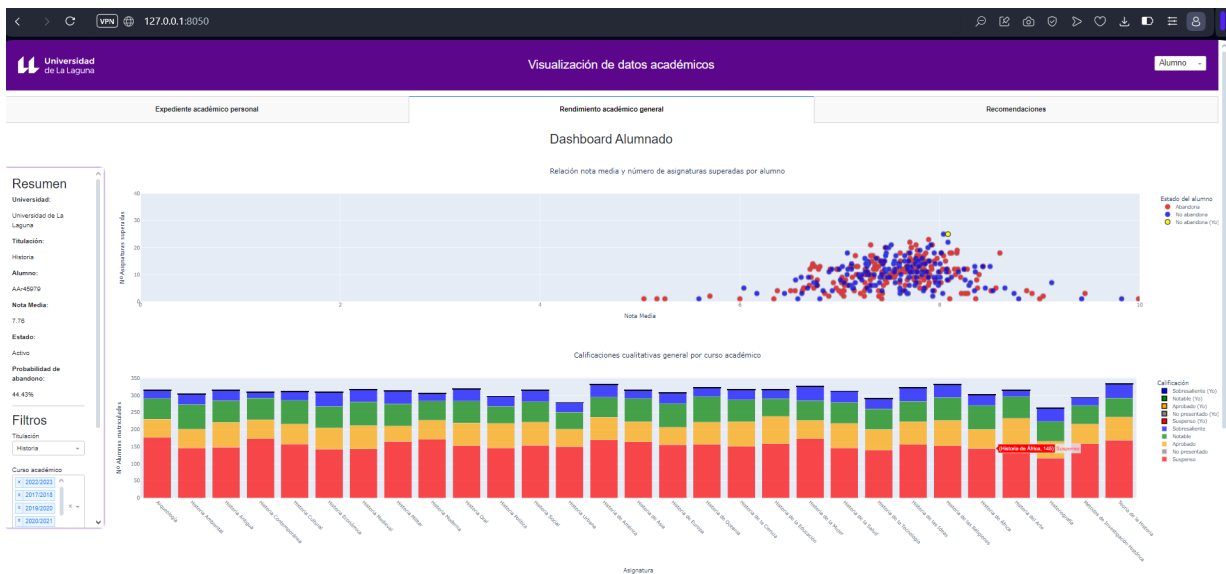


Figura 5.6: Perfil alumnado. Pestaña rendimiento académico general.

Las representaciones gráficas que se visualizan son las siguientes:

- **Relación nota media con asignaturas superadas por alumno:** Esta representación se visualiza a través de un gráfico de dispersión. Cada punto representa a un alumno. El color rojo indica que el alumno ha abandonado la titulación, el azul indica que el alumno no abandona la titulación y el color amarillo representa al propio alumno. Se utiliza esta representación para observar la relación que existe entre las asignaturas superadas con la nota media. De esta forma, el alumno puede ver los tipos de puntos que le rodean, indicando el comportamiento común de los alumnos (Figura 5.7).

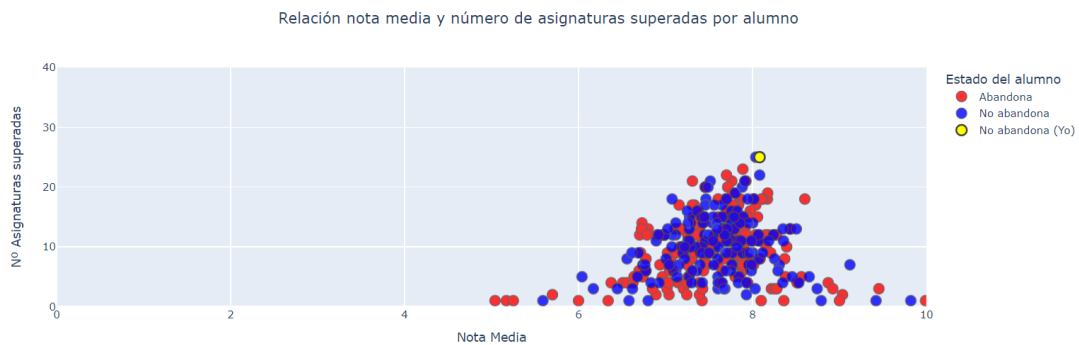


Figura 5.7: Gráfica. Relación nota media y asignaturas superadas.

- **Calificaciones cualitativas:** Representa el número de alumnos según la calificación cualitativa obtenida por asignatura. El alumno puede ver su calificación en la zona superior de cada barra con un tono de color oscuro para diferenciar con los resultados de los compañeros. Los alumnos comparan sus calificaciones con las de sus compañeros, lo cual puede motivarles a mejorar y proporcionar una visión clara de su rendimiento (Figura 5.8).

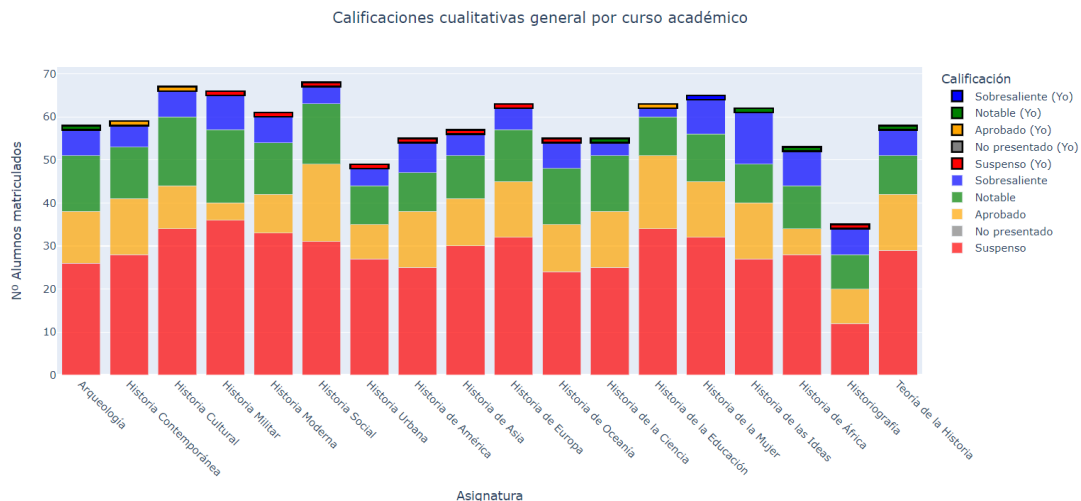


Figura 5.8: Gráfica. Calificaciones cualitativas general.

- **Relación calificaciones del alumno con la nota media general:** Para mejorar la visualización de la comparativa, se optó por utilizar un gráfico de barras agrupadas. El color azul indica la nota media del alumno y el color gris representa la nota media general por asignatura. Esta representación permite al alumno ver en qué asignaturas se encuentra por encima o por debajo de la media de la clase, lo cual puede ser un indicador para enfocar sus esfuerzos de estudio (Figura 5.9).

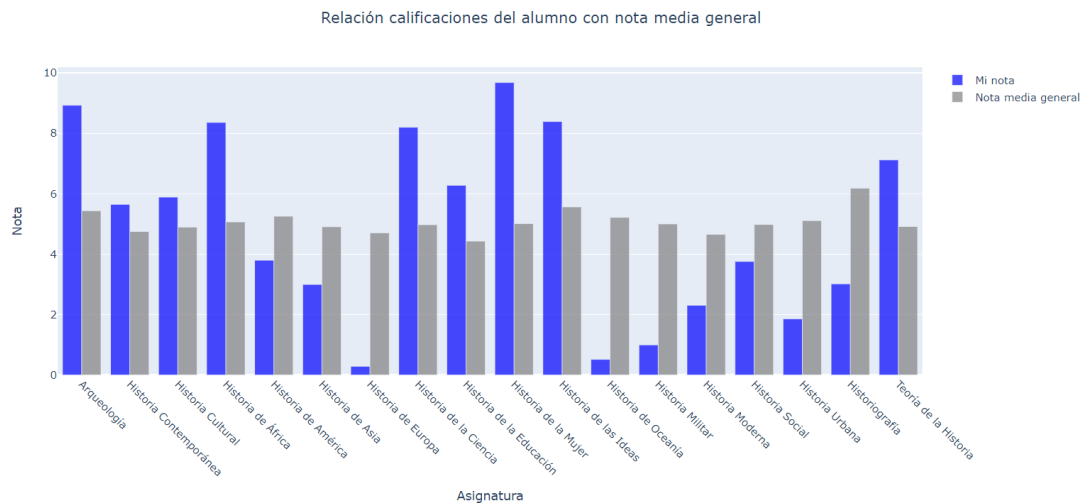


Figura 5.9: Gráfica. Relación calificaciones del alumno con nota media general.

- **Recomendaciones:** Esta pestaña contiene información relevante sobre el abandono académico. El enfoque principal es orientar al alumnado sobre esta problemática mediante información de interés (Figura 5.10).

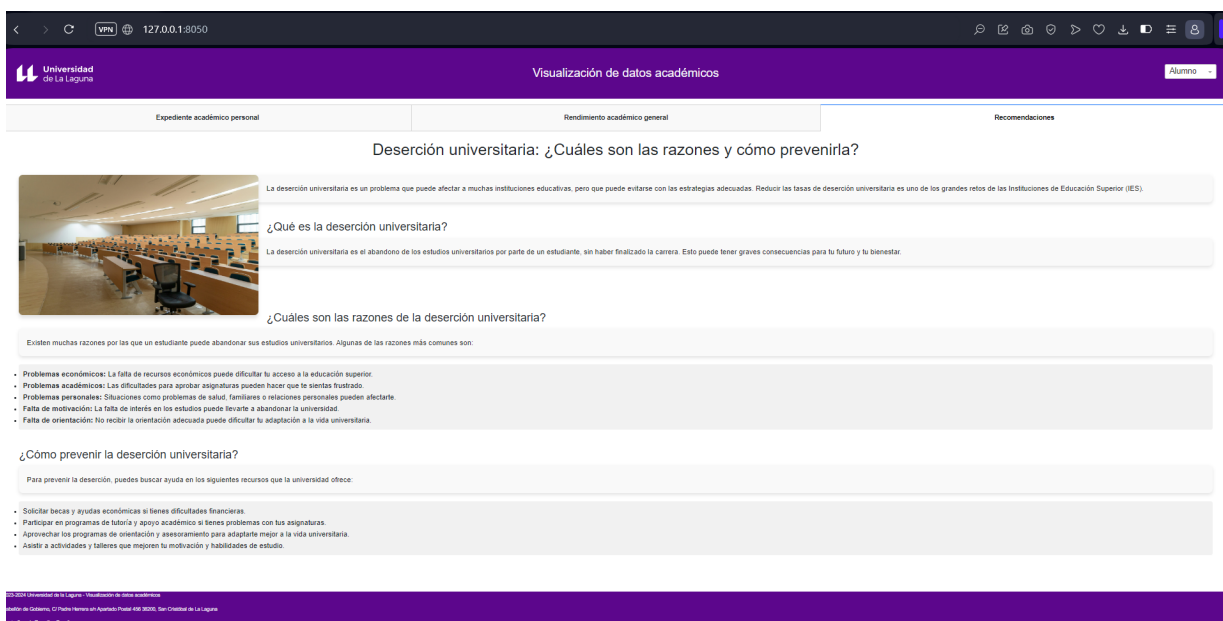


Figura 5.10: Perfil alumnado. Pestaña recomendaciones.

5.1.2. Perfil docente

En el perfil docente también se plantearon cambios y funcionalidades adicionales, con tres divisiones similares a las del perfil alumnado. Las representaciones se dividen de la siguiente forma:

- **Rendimiento académico personal:** Esta sección representa los resultados académicos de los alumnos en aquellas asignaturas que imparte el docente. Se perseguía la idea de que el docente pueda ver fácilmente la evolución por curso académico (Figura 5.11). Al igual que el perfil alumnado, se presenta un panel con un resumen con información: universidad, titulación y nombre del docente. Además los filtros que interactúan con los gráficos en este caso son: titulación, curso académico, asignaturas.

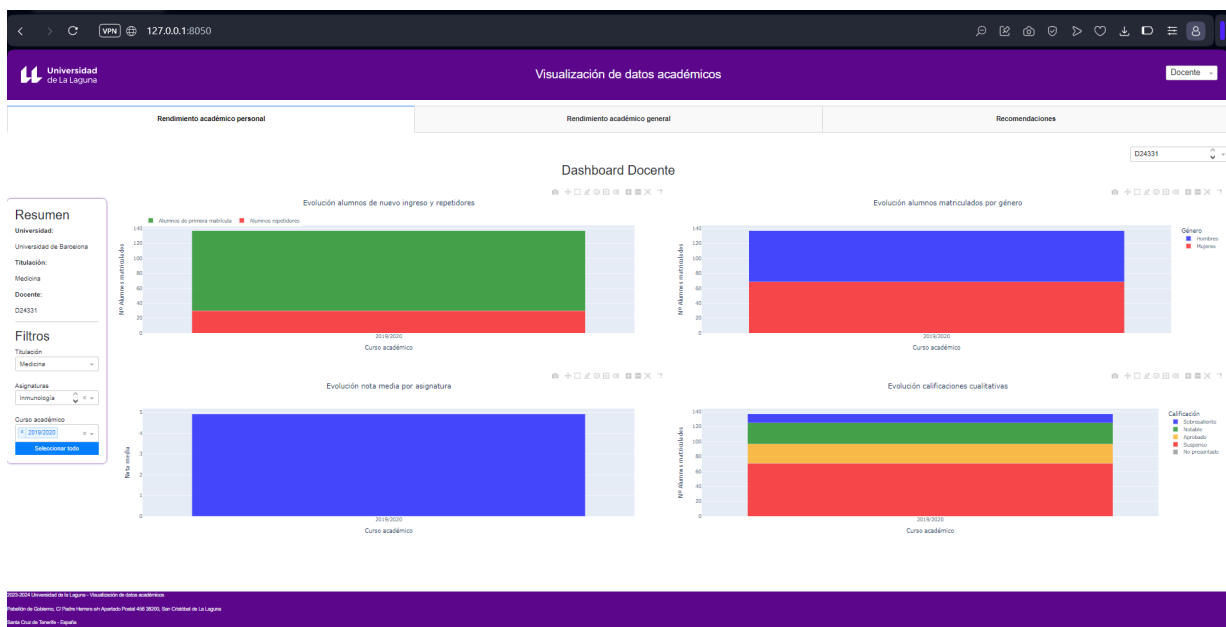


Figura 5.11: Perfil docente. Pestaña rendimiento académico personal.

Las representaciones gráficas que se visualizan son las siguientes:

- **Evolución alumnos de nuevo ingreso y repetidores:** La gráfica de barras apiladas se eligió como la representación más adecuada para mostrar la evolución por curso académico de la cantidad de alumnos de nuevo ingreso y repetidores por asignatura, ya que permitía obtener una mejor visualización de los resultados. Gracias a esta información el docente puede ver en que asignaturas hay problemas de superación (Figura 5.12).
- **Evolución alumnos matriculados por género:** Se utiliza el gráfico de barras apiladas para representar la cantidad de alumnos que hay según el género por curso académico. Esta visualización ayuda a identificar posibles desigualdades de género en la matriculación (Figura 5.13).

- **Evolución calificaciones cualitativas:** También se representan las calificaciones cualitativas de los alumnos por curso académico con un gráfico de barras apiladas. Permite a los docentes observar la distribución de las calificaciones y evaluar el rendimiento global de sus estudiantes (Figura 5.14).
- **Evolución nota media por asignaturas:** Esta representación se visualiza con un gráfico de barras para ver la evolución de la media de los alumnos por curso académico (Figura 5.15).

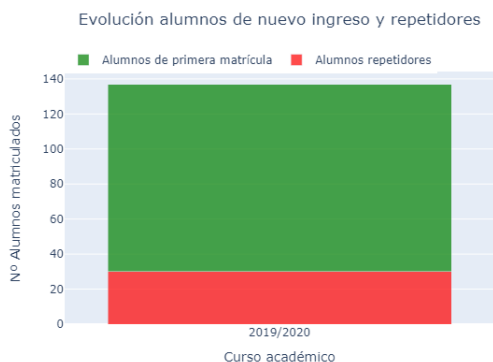


Figura 5.12: Gráfica. Evolución alumnos de nuevo ingreso y repetidores



Figura 5.13: Gráfica. Evolución alumnos matriculados por género

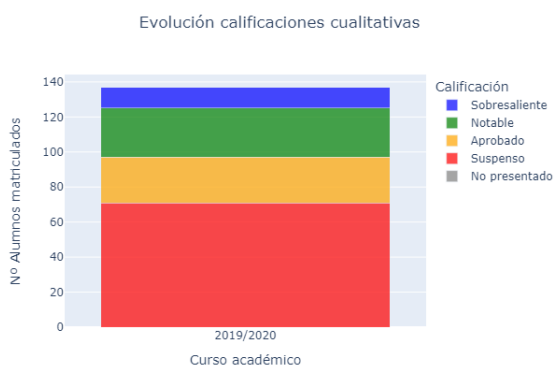


Figura 5.14: Gráfica. Evolución calificaciones cualitativas

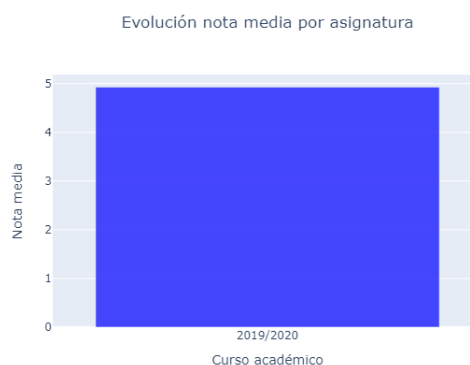


Figura 5.15: Gráfica. Evolución nota media por asignatura

- **Rendimiento académico general:** Permite al docente comparar las calificaciones de las asignaturas que imparte con el resto de asignaturas de la titulación, con el propósito de obtener una perspectiva más amplia y evaluar el estado de sus alumnos con respecto a otras asignaturas (Figura 5.16).

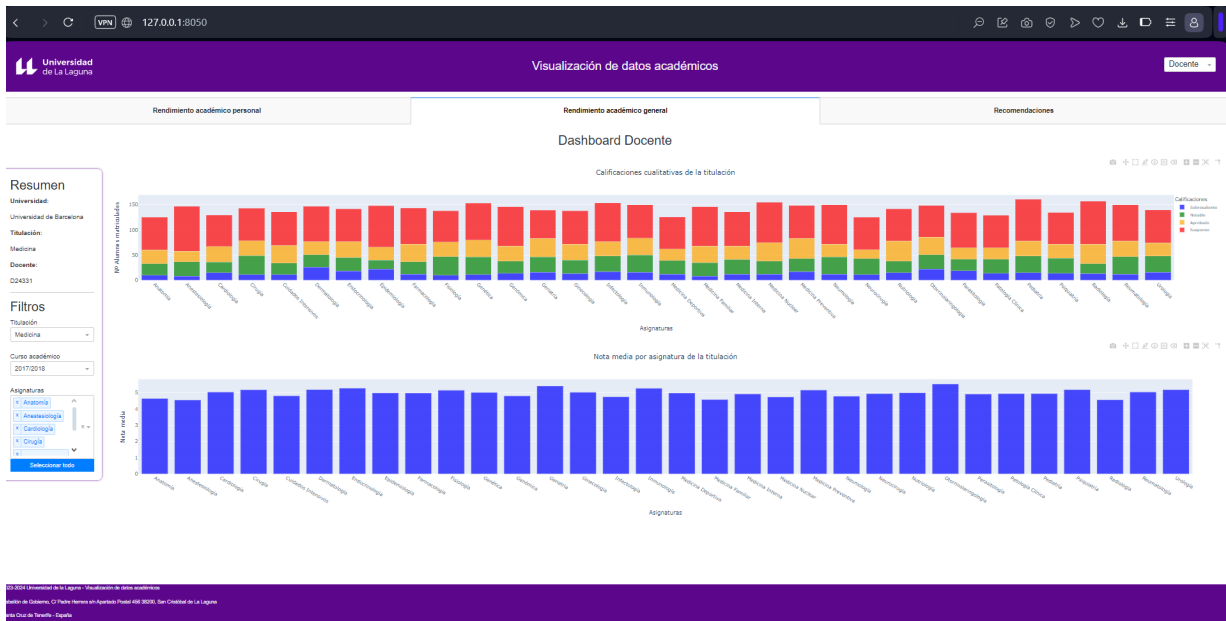


Figura 5.16: Perfil docente. Pestaña rendimiento académico general.

- **Calificaciones cualitativas de la titulación:** Se utiliza un gráfico de barras apiladas para visualizar la cantidad de alumnos según el tipo de calificación obtenida por curso académico y asignatura. Esta gráfica permite ver la distribución de calificaciones cualitativas en las diferentes asignaturas, lo que puede indicar el nivel de dificultad y la efectividad de la enseñanza impartida (Figura 5.17).

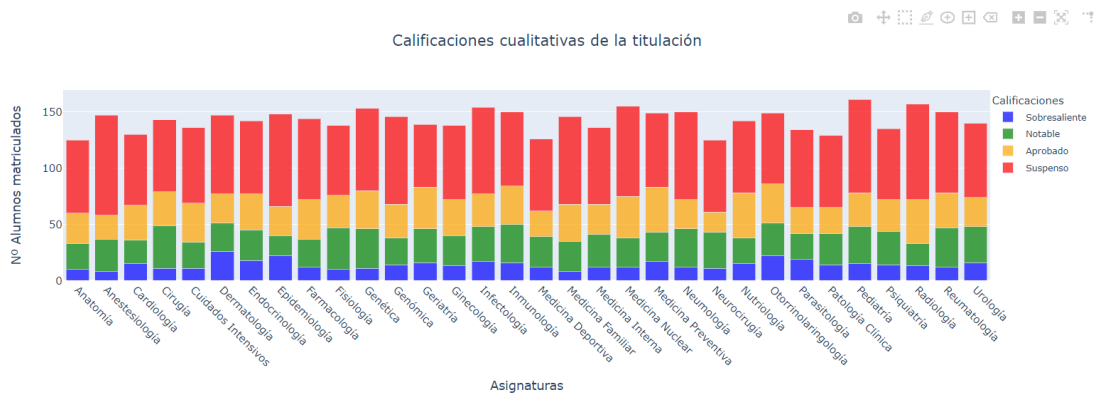


Figura 5.17: Gráfica. Calificaciones cualitativas de la titulación.

- **Nota media por asignatura de la titulación:** Para representar la nota media por asignatura según el curso académico se utiliza un gráfico de barras. Esta visualización permite a los docentes identificar tendencias en las calificaciones medias de las diferentes asignaturas (Figura 5.18).

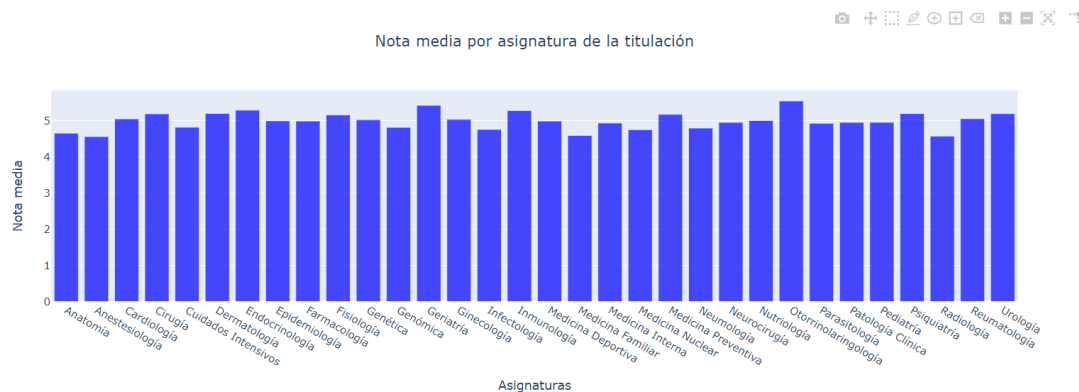


Figura 5.18: Gráfica. Nota media por asignaturas de la titulación.

- **Recomendaciones:** De manera análoga al perfil alumnado, la pestaña recomendaciones contiene información útil relativa a la deserción académica. En este caso, la información está enfocada hacia los docentes (Figura 5.19).

Deserción universitaria: ¿Cómo pueden ayudar los docentes a prevenirla?

La deserción universitaria es un problema serio que afecta a muchas instituciones educativas. Los docentes juegan un papel crucial en la identificación y prevención de este fenómeno.

¿Qué es la deserción universitaria?

La deserción universitaria es el abandono de los estudios universitarios por parte de un estudiante, sin haber finalizado la carrera. Este problema puede tener graves consecuencias tanto para los estudiantes como para la institución.

¿Cómo pueden los docentes ayudar a prevenir la deserción universitaria?

- Identificar y apoyar a los estudiantes con dificultades académicas proporcionando tutorías y asesorías.
- Fomentar un ambiente de clase inclusivo y motivador que promueva la participación y el interés por el aprendizaje.
- Mantener una comunicación abierta y constante con los estudiantes para entender sus problemas y necesidades.
- Colaborar con los servicios de apoyo estudiantil para proporcionar la ayuda necesaria en caso de problemas personales o económicos.
- Participar en programas de formación continua para mejorar las estrategias pedagógicas y adaptarse a las necesidades cambiantes de los estudiantes.

Figura 5.19: Perfil docente. Pestaña recomendaciones.

5.1.3. Perfil gestor

En el perfil gestor se realizaron cuatro divisiones. Cada gestor puede visualizar los datos según la universidad que tenga asignada (Figura 5.20). Las representaciones se dividen de la siguiente forma:

- **Indicadores académicos:** Los indicadores representan información relativa a la clasificación de alumnos según características por cada una de las titulaciones. El panel de resumen contiene la siguiente información: universidad, nombre del gestor y número de alumnos matriculados. Con respecto a los filtros que interactúan con los gráficos, se utilizan: curso académico y titulaciones. Se dispone de la siguiente información:

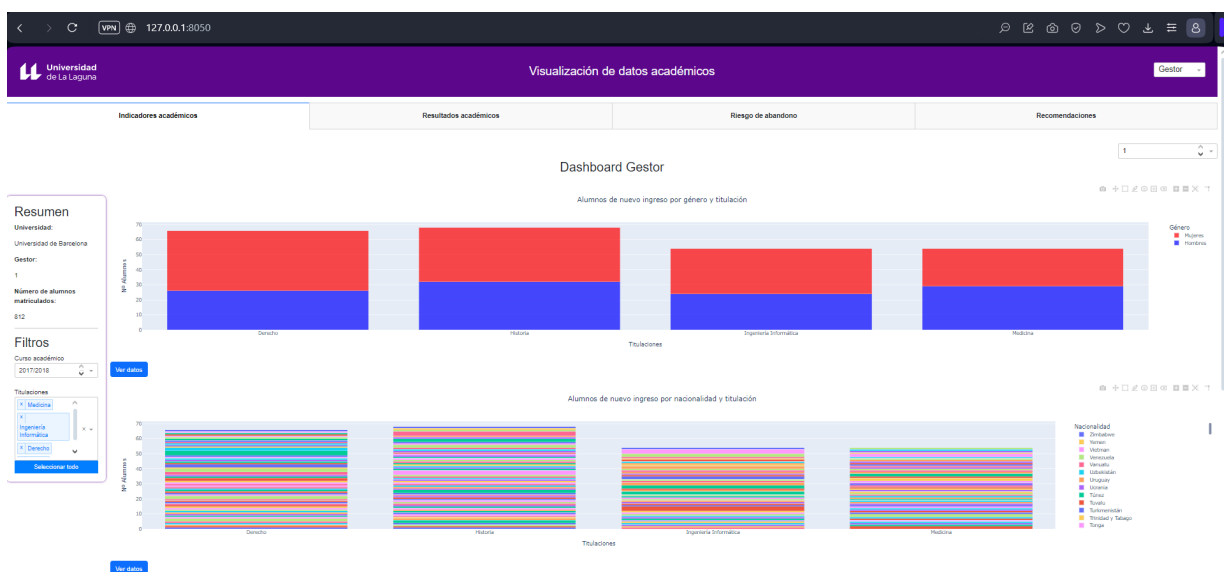


Figura 5.20: Perfil gestor. Pestaña indicadores.

- Alumnos de nuevo ingreso por género y titulación (Figura 5.21).
- Alumnos de nuevo ingreso por nacionalidad y titulación (Figura 5.22).
- Alumnos egresados por género y titulación (Figura 5.23).
- Alumnos egresados por nacionalidad y titulación (Figura 5.24).

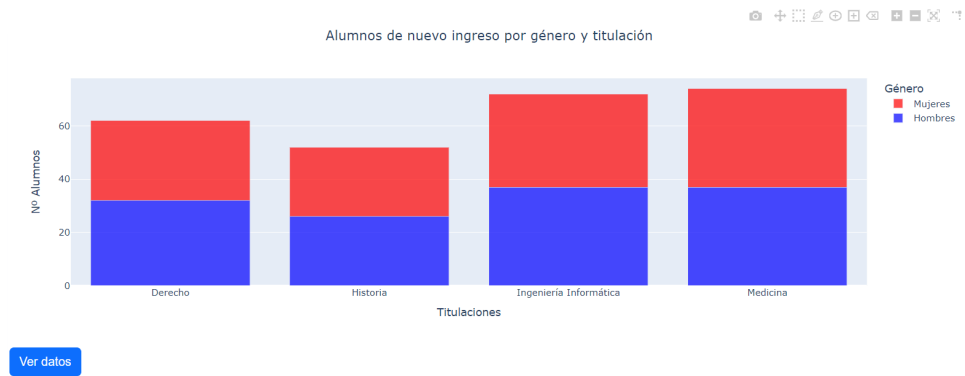


Figura 5.21: Gráfica. Alumnos de nuevo ingreso por género y titulación.

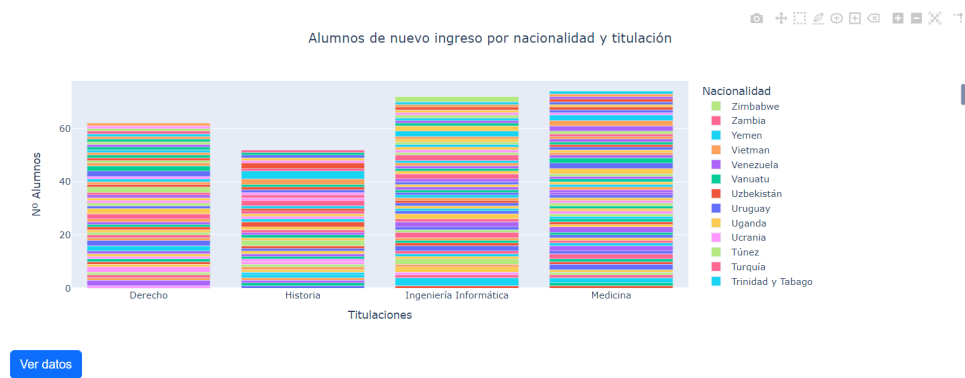


Figura 5.22: Gráfica. Alumnos de nuevo ingreso por nacionalidad y titulación.

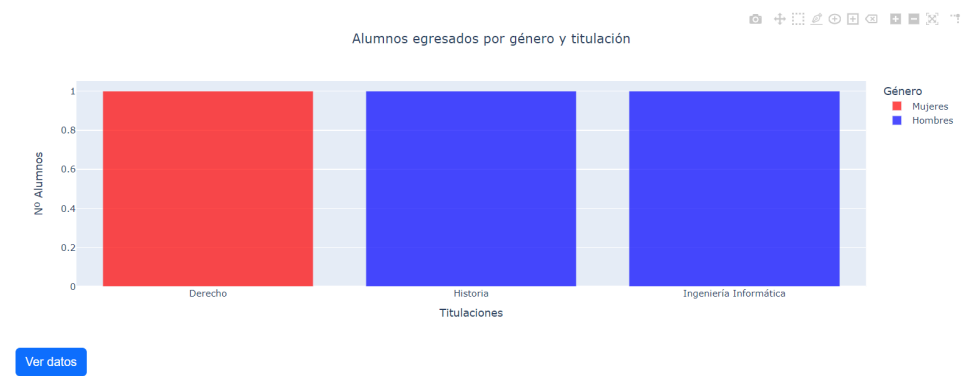


Figura 5.23: Gráfica. Alumnos egresados por género y titulación.



Figura 5.24: Gráfica. Alumnos egresados por nacionalidad y titulación.

Las representaciones de este cuadro de mando se realizan con gráficos de barras apiladas para poder visualizar de mejor manera los resultados. Esta información proporcionada es útil para identificar posibles desigualdades de género en la admisión de estudiantes y evaluar la efectividad de las políticas de inclusión. Así como para comprender la diversidad y conseguir adaptar los programas académicos y servicios de apoyo a las necesidades de estudiantes de diferentes orígenes culturales y nacionales.

- **Resultados académicos:** En esta sección se muestran los datos relacionados con las calificaciones por titulaciones. En este caso concreto, no se hace uso de filtros al no ser necesario (Figura 5.25).

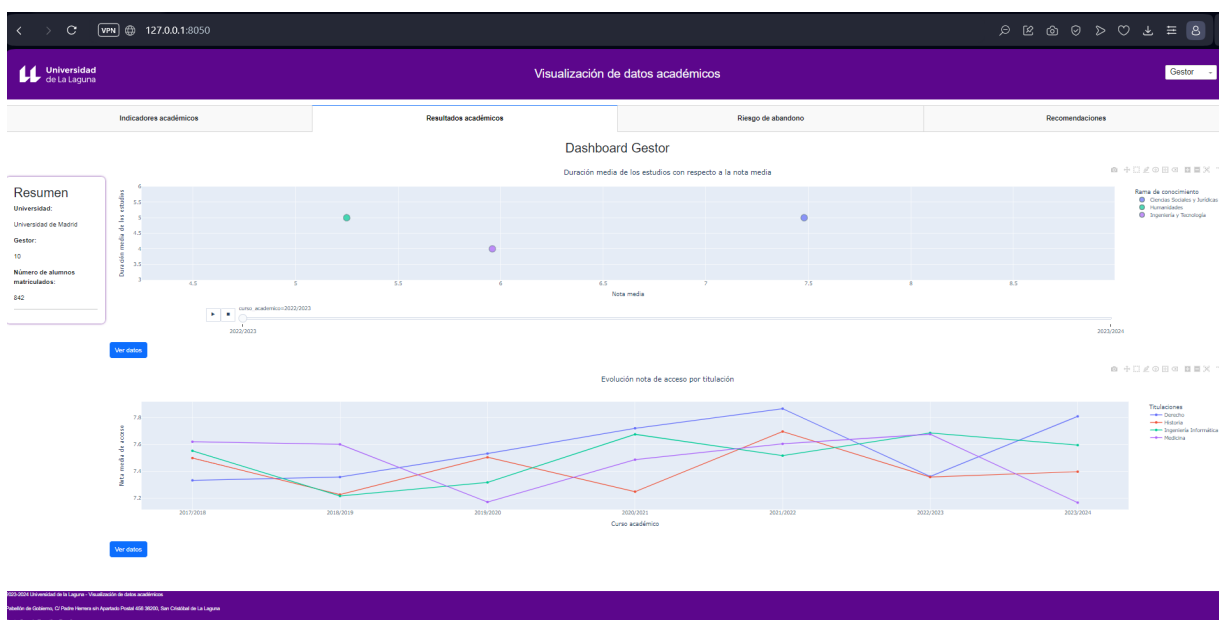


Figura 5.25: Perfil gestor. Pestaña resultados.

- **Duración media de los estudios con respecto a la nota media:** Se utiliza un gráfico de burbujas. Cada burbuja representa una titulación, el color representa la rama del conocimiento, el tamaño representa el número de alumnos egresados, el eje X representa la nota media y el eje Y representa la duración media de los estudios. Además, se dispone de un control deslizante con los cursos académicos para representar la evolución temporal (Figura 5.26).



Figura 5.26: Gráfica. Duración media de los estudios con respecto a la nota media.

- **Evolución nota de acceso por titulación:** Se utiliza un gráfico de líneas para visualizar mejor la evolución en el tiempo de las notas de acceso a la universidad por titulación. Esta representación permite ver cómo han variado los requisitos de ingreso, lo cual puede reflejar cambios en la demanda y la competitividad de las titulaciones (Figura 5.27).

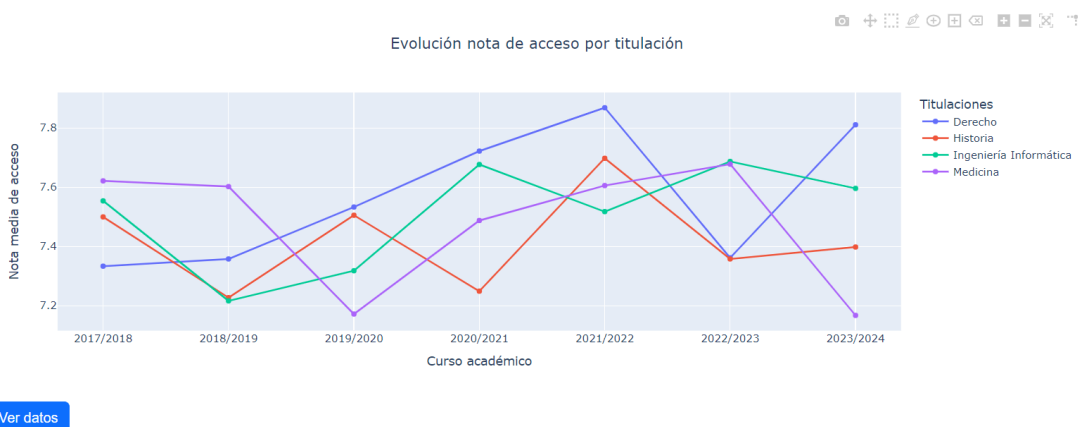


Figura 5.27: Gráfica. Evolución nota de acceso por titulación.

- **Riesgo de abandono:** Esta pestaña contiene los elementos que permiten tener información sobre la deserción académica (Figura 5.28). En este caso sólo se utiliza el filtro de curso académico para actualizar los gráficos.

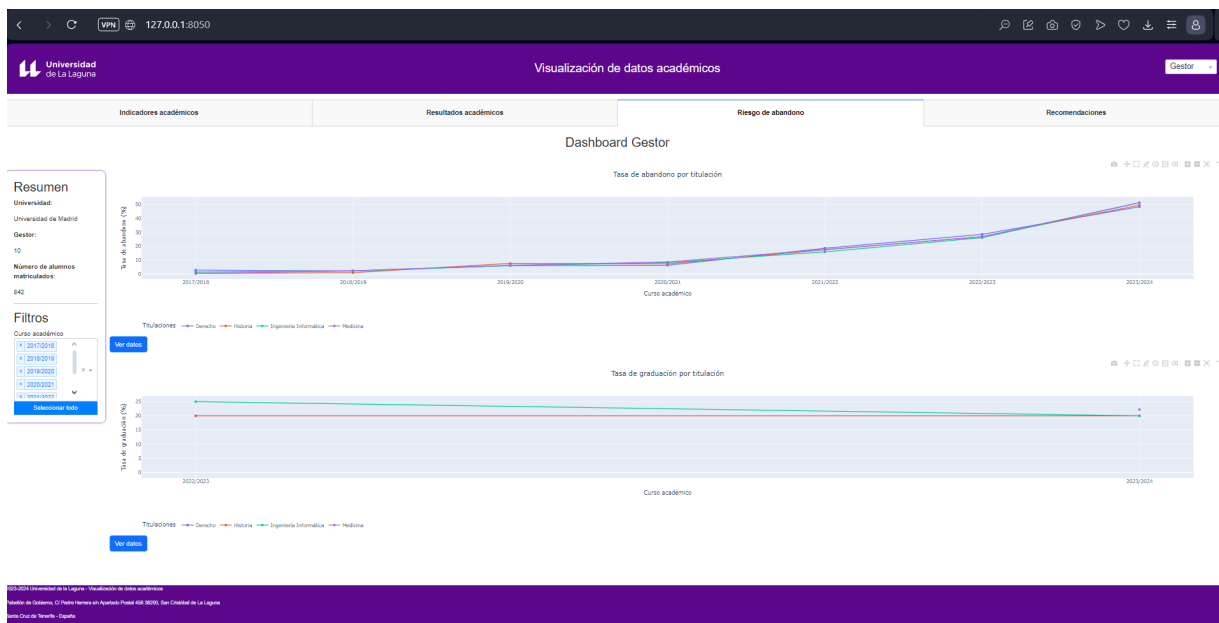


Figura 5.28: Perfil gestor. Pestaña riesgo de abandono.

- **Tasa de abandono por titulación:** A través de un gráfico de líneas se representa la cantidad de alumnos que han abandonado la titulación en cada curso académico. Esta métrica permite entender la magnitud del problema de deserción y tomar medidas específicas para apoyar a los estudiantes en riesgo de abandono (Figura 5.29). Se calcula de la siguiente manera:

$$\text{Tasa de abandono} = \frac{\text{Número de alumnos que abandonan}}{\text{Número de matriculados}} * 100$$

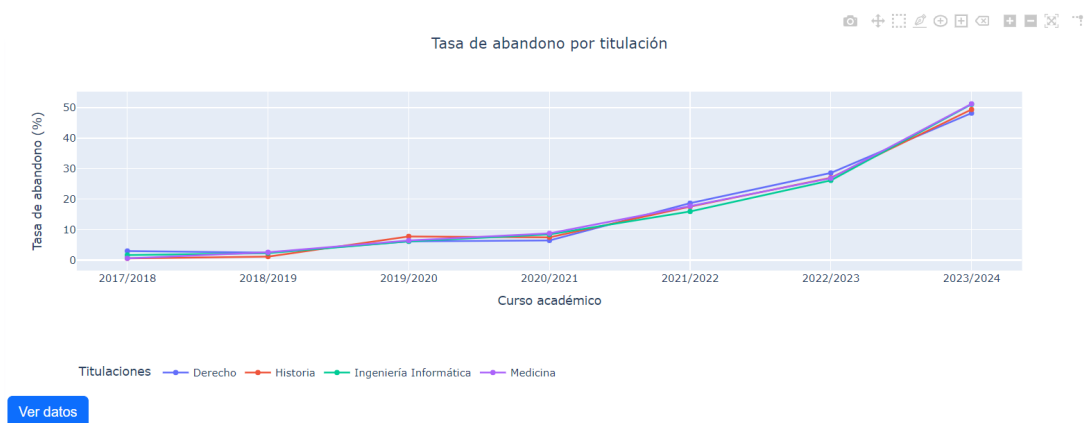


Figura 5.29: Gráfica. Tasa de abandono por titulación.

- **Tasa de graduación por titulación:** Representa la tasa de graduación de los alumnos por titulación mediante un gráfico de líneas. Esta representación es esencial para evaluar la efectividad de los programas académicos en términos de completar la titulación (Figura 5.30).

$$\text{Tasa de graduación} = \frac{\text{Número de alumnos que finalizan la titulación}}{\text{Número de matriculados}} * 100$$

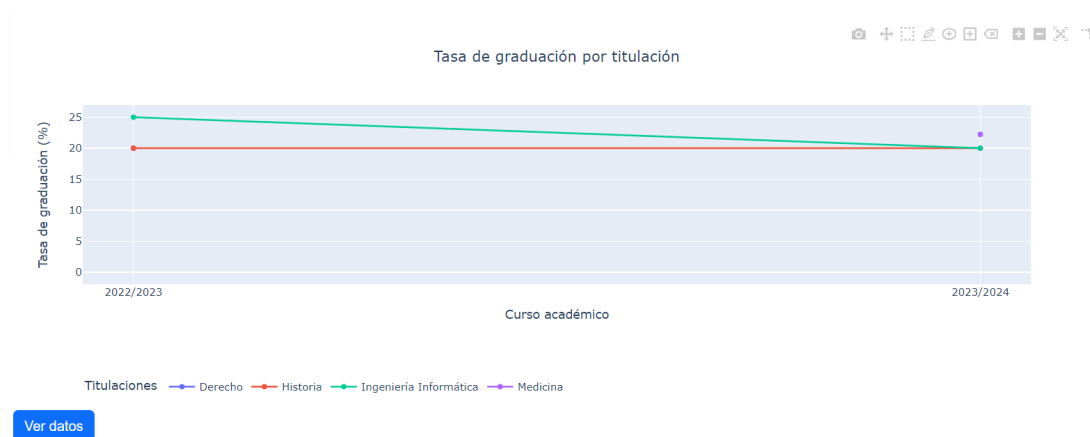


Figura 5.30: Gráfica. Tasa de graduación por titulación.

- **Recomendaciones:** Esta sección comparte el mismo objetivo que los perfiles anteriores, información relativa a la deserción académica. En este caso, la información está enfocada hacia los gestores de las universidades (Figura 5.31).

Deserción universitaria

La deserción universitaria es un desafío significativo para las instituciones de educación superior. Los gestores tienen la responsabilidad de desarrollar e implementar políticas efectivas para reducirla.

¿Qué es la deserción universitaria?

La deserción universitaria se refiere al abandono de los estudios por parte de los estudiantes antes de completar su formación. Este fenómeno tiene implicaciones negativas para la universidad y para la sociedad.

Estrategias para prevenir la deserción universitaria

- Desarrollar programas de becas y ayudas financieras para apoyar a los estudiantes con dificultades económicas.
- Implementar sistemas de tutoría y mentoría para ofrecer apoyo académico y emocional a los estudiantes.
- Establecer programas de orientación y adaptación para nuevos estudiantes, facilitando su integración en la vida universitaria.
- Promover la formación continua del personal docente en metodologías pedagógicas innovadoras.
- Utilizar sistemas de seguimiento y análisis de datos para identificar a estudiantes en riesgo y actuar de manera preventiva.
- Fomentar un clima institucional inclusivo y de apoyo que motive a los estudiantes a continuar con sus estudios.

Figura 5.31: Perfil gestor. Pestaña recomendaciones.

En los perfiles docente y gestor, cada gráfico dispone de una herramienta con varias funcionalidades para mejorar la experiencia del usuario. Algunas de las funcionalidades son: capturar una imagen del gráfico y descargarla, mover el gráfico, selección por áreas, dibujar líneas, cuadrados, círculos, ampliar, reducir, autoescala y una guía en los ejes (Figura 5.32).

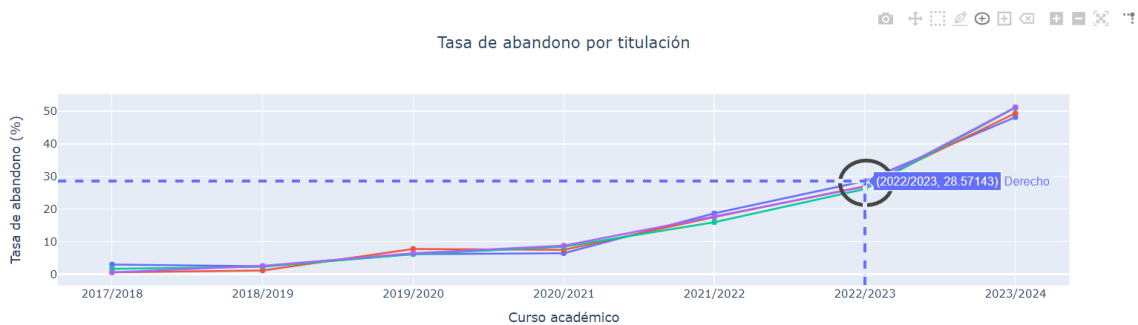


Figura 5.32: Ejemplo herramientas de gráficos.

En el perfil de gestor, está disponible una funcionalidad adicional: un botón ubicado en la esquina inferior izquierda de cada gráfico. Esta funcionalidad permite visualizar los datos utilizados en forma de tabla y descargarlos en formato CSV (Figura 5.33).

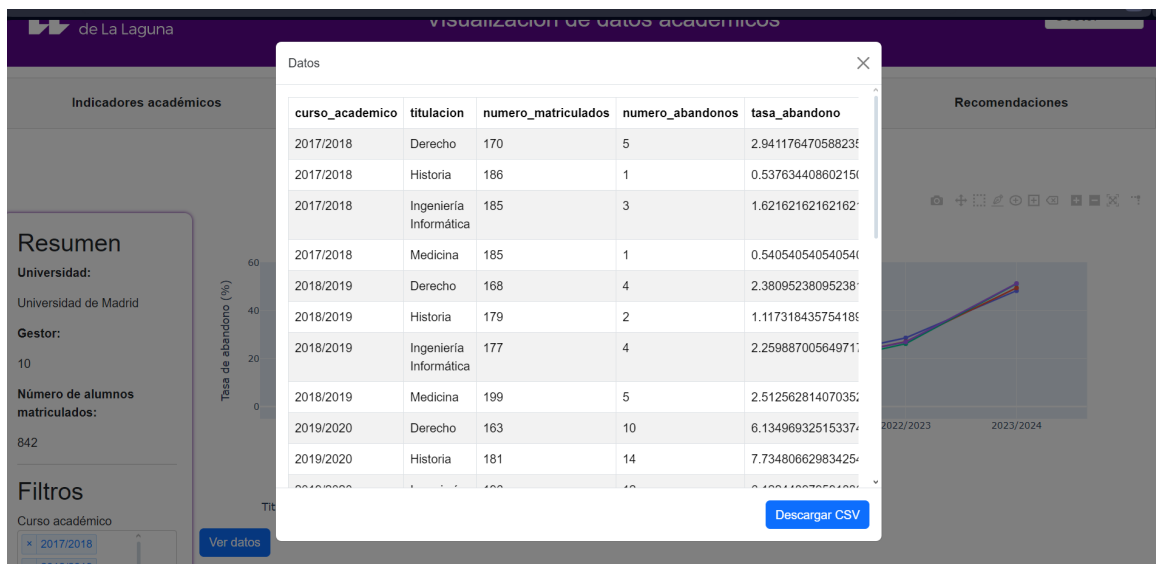


Figura 5.33: Tabla de datos.

5.2. Resultados del modelo predictivo

Para medir el rendimiento y la efectividad de un modelo predictivo en base a cómo se comporta con datos de prueba se utilizan unas métricas de evaluación, las métricas utilizadas en el proyecto fueron las siguientes:

1. **Matriz de confusión:** Proporciona un resumen del desempeño del modelo al mostrar la cantidad de predicciones correctas e incorrectas (Figura 5.34).

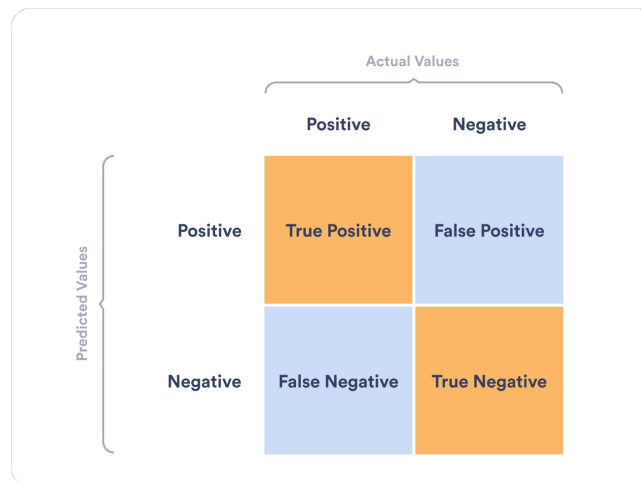


Figura 5.34: Matriz de confusión.

- **Verdaderos positivos (TP):** Instancias positivas que el modelo predijo correctamente como positivas.
 - **Verdaderos negativos (TN):** Instancias negativas que el modelo predijo correctamente como negativas.
 - **Falsos positivos (FP):** Instancias negativas que el modelo predijo incorrectamente como positivas.
 - **Falsos negativos (FN):** Instancias positivas que el modelo predijo incorrectamente como negativas.
2. **Accuracy (Exactitud):** Es la proporción de predicciones correctas (tanto positivas como negativas) realizadas por el modelo respecto al total de predicciones realizadas, indica la proporción de aciertos totales del modelo, se calcula como:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

3. **Precision (Precisión):** Indica la proporción de predicciones positivas que fueron correctas. Es útil cuando el costo de los falsos positivos es alto. Una alta precisión significa que las predicciones positivas del modelo son confiables. Se calcula como:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

4. **Recall (Sensibilidad):** Mide la proporción de instancias positivas que fueron correctamente identificadas por el modelo. Es útil cuando es importante identificar todas las instancias positivas, sin importar los falsos positivos. Se calcula como:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

5. **F1 Score:** Es la media armónica de precisión y recall, proporcionando un balance entre ambas métricas. Es útil cuando hay un desbalance significativo entre las clases objetivo. Un alto F1 score indica un buen balance entre precisión y recall. Se calcula como:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

6. **ROC AUC (Área bajo la Curva ROC):** La curva ROC es una representación gráfica de la sensibilidad (recall) frente a la tasa de falsos positivos. El área bajo esta curva proporciona una medida del rendimiento discriminativo del modelo:

- ROC AUC = 0.5: Clasificación aleatoria
- ROC AUC >0.5: Mejor que clasificación aleatoria
- ROC AUC = 1: Clasificación perfecta

Como se mencionó en el capítulo 4.4. Tras realizar el entrenamiento de cada modelo se obtuvieron los siguientes resultados:

Modelo	Accuracy	Precision	Recall	F1 Score	ROC AUC
RandomForest	0.49	0.49	0.49	0.49	0.49
LogisticRegression	0.49	0.50	0.45	0.47	0.49
KNeighbors	0.51	0.51	0.50	0.51	0.51
AdaBoost	0.48	0.38	0.04	0.07	0.49
XGBoost	0.48	0.49	0.50	0.49	0.48
LightGBM	0.51	0.52	0.41	0.46	0.51
VotingClassifier	0.52	0.52	0.51	0.52	0.52

Tabla 5.1: Resultados de los modelos

Modelo	TP	FP	FN	TN
RandomForest	241	254	257	248
LogisticRegression	268	227	280	225
KNeighbors	255	240	252	253
AdaBoost	461	34	484	21
XGBoost	227	268	251	254
LightGBM	301	194	297	208
VotingClassifier	259	236	247	258

Tabla 5.2: Matrices de confusión de los modelos

Los modelos con mejores resultados son: KNeighbors, RandomForest y XGBoost. Se seleccionaron los mejores modelos basados en la métrica F1 Score, ya que proporciona un buen equilibrio entre precisión y sensibilidad para crear el VotingClassifier. Como se observa en la tabla de resultados, combina las fortalezas de los mejores algoritmos para obtener mayor robustez.

Es importante tener en cuenta que los datos utilizados en el modelo son completamente aleatorios. Esto puede afectar los resultados de las métricas de evaluación del modelo. Los datos aleatorios suelen no contener patrones significativos, lo que puede llevar a resultados inesperados en términos de precisión, sensibilidad y otras métricas.

Para obtener información significativa y una evaluación más precisa del rendimiento del modelo, se deben utilizar datos reales que representen fielmente el problema que se desea resolver.

El resultado de la probabilidad de abandono se ve reflejada en un indicador en el dentro del panel de resumen ubicado en el perfil del propio alumno, como ve en la Figura 5.35.

Resumen	
Universidad:	Universidad de La Laguna
Titulación:	Historia
Alumno:	AAr45979
Nota Media:	7.76
Estado:	Activo
Probabilidad de abandono:	44.43%

Figura 5.35: Resultado de probabilidad de abandono. Perfil alumnado

5.3. Divulgación de la herramienta

Los días 6 y 7 de junio se celebraron las I Jornadas sobre Abandono de los Estudios Universitarios, organizadas por el Grupo Universitario de Formación y Orientación Integrada (GUFOI).

Estas jornadas están vinculadas al proyecto I+D+i titulado *“Análisis de los factores explicativos del abandono de los estudios universitarios y acciones estratégicas para su mejora y prevención”*, mencionado anteriormente. Con el proyecto llegando a su fin este año, se han organizado estas jornadas para presentar los resultados finales y discutir la implementación de soluciones.

El evento contó con la participación de los miembros del equipo de trabajo del proyecto, así como con los vicerrectores de calidad y docencia de las universidades participantes, representantes de la Consejería de Universidades y la Consejería de Educación del Gobierno de Canarias. Las jornadas también estuvieron abiertas a profesionales relacionados con la educación y la orientación al alumnado de secundaria, bachillerato y universidades [27].

Además de las presentaciones formales, las jornadas incluyeron mesas redondas donde se discutieron diversas estrategias para abordar el abandono universitario. Durante el evento, en la última mesa redonda, se presentó el proyecto de Trabajo de Fin de Grado en cuestión. Este proyecto fue divulgado como una herramienta complementaria, aportando valor a las investigaciones realizadas por estudiantes de Máster y Doctorado en el área educativa [26]. Esta solución informática, ha servido como un recurso práctico para identificar patrones y factores que contribuyen al abandono universitario, ofreciendo una amplia perspectiva de los resultados.

Capítulo 6

Conclusiones y líneas futuras

6.0.1. Conclusiones

Al terminar este proyecto de fin de grado, se ha logrado cumplir con los objetivos propuestos inicialmente y ofrecer soluciones valiosas al problema. Después de comparar varias herramientas de inteligencia empresarial, se eligió la que mejor se adaptaba al desarrollo de aplicaciones analíticas e interactivas. Esta elección optimizó el proceso de desarrollo y aseguró una implementación más eficiente de las soluciones propuestas, permitiendo además su libre uso y distribución

Los cuadros de mando interactivos desarrollados han demostrado ser herramientas eficaces para visualizar datos académicos, permitiendo a los usuarios identificar patrones y tendencias de manera intuitiva y dinámica. Este enfoque puede permitir una mejor comprensión y análisis del rendimiento estudiantil.

Un aspecto crucial ha sido la evaluación de resultados mediante la generación de datos sintéticos. Esta medida ha garantizado la evaluación de los resultados obtenidos y el funcionamiento de la herramienta.

La integración de un modelo predictivo utilizando técnicas de Machine Learning ha permitido realizar una posible estimación de la probabilidad de abandono académico de cada estudiante, basándose en datos históricos y actuales. Las predicciones facilitan la identificación oportuna de los estudiantes en riesgo y la implementación de medidas preventivas adecuadas.

Para finalizar, el desarrollo de esta aplicación abre el camino a poder demostrar si el uso eficiente de herramientas de Business Intelligence y técnicas de Machine Learning puede mejorar significativamente la gestión del rendimiento estudiantil, siendo este el objetivo principal del proyecto.

6.0.2. Líneas futuras

En esta sección se presentan las funcionalidades que podrían mejorar y ampliar la herramienta en un futuro.

- **Sistema de registro de usuarios:** Un aspecto importante es la implementación de un sistema de registro de usuarios utilizando las credenciales universitarias. Esto permitiría un acceso personalizado para cada usuario, asegurando la correcta gestión de permisos y niveles de acceso según el rol.
- **Conexión con la base de datos real:** Utilizar la base de datos real del proyecto ABACU para evaluar el funcionamiento objetivo de la herramienta.
- **Despliegue:** Planificar y ejecutar el despliegue de la herramienta en el entorno de producción de la universidad. Este proceso debe incluir la configuración del servidor, la optimización del rendimiento y la integración del modelo de aprendizaje automático.
- **Contenerización:** Utilizar Docker, para crear versiones portátiles y fácilmente desplegables de la herramienta para facilitar la gestión de dependencias y la ejecución en diferentes entornos.
- **Búsqueda y visualización individualizada de alumnos:** Se podría agregar una vista específica para docentes y gestores que permita buscar a un alumno en particular y visualizar sus resultados académicos detallados. Esta funcionalidad facilitaría el seguimiento individualizado y la intervención cuando sea necesario.
- **Alertas y notificaciones:** Implementar un sistema de alertas tempranas que notifique a los profesores sobre los alumnos que se encuentren en un estado de riesgo de abandono.
- **Transformación en una plataforma integral de apoyo educativo:** La herramienta podría evolucionar hacia una plataforma integral de apoyo educativo que también ofrezca recursos y recomendaciones personalizadas. Esta plataforma podría incluir:
 - **Recursos educativos personalizados:** Ofrecer materiales y recursos educativos adaptados a las necesidades individuales de los alumnos, basados en sus áreas de mejora y fortalezas.
 - **Programas de mentoría:** Conectar a los alumnos en riesgo con mentores que puedan guiarlos y apoyarlos en su desarrollo académico y personal.

Capítulo 7

Summary and Conclusions

At the end of this final degree project, the initially proposed objectives have been achieved and valuable solutions to the problem have been offered. After comparing several business intelligence tools, the one best suited to the development of analytical and interactive applications was chosen. This choice optimized the development process and ensured a more efficient implementation of the proposed solutions, while allowing its free use and distribution.

The integration of a predictive model using Machine Learning techniques has enabled a possible estimation of the probability of academic dropout for each student, based on historical and current data. The predictions facilitate the timely identification of students at risk and the implementation of appropriate preventive measures.

A crucial aspect has been the evaluation of results through the generation of synthetic data. This measure has ensured the evaluation of the results obtained and the performance of the tool.

The integration of a predictive model using Machine Learning techniques has allowed to accurately estimate the probability of academic dropout for each student, based on historical and current data. The predictions facilitate the timely identification of at-risk students and the implementation of appropriate preventive measures.

Finally, the development of this application opens the way to demonstrate whether the efficient use of Business Intelligence tools and Machine Learning techniques can significantly improve student performance management, which was the main objective of the project.

Capítulo 8

Presupuesto

En este apartado se detalla los costes asociados al proyecto:

Concepto	Detalle	Coste
Desarrollo de Software		5.100€
Análisis	40 horas x 17 €/hora	680€
Planificación	40 horas x 17 €/hora	680€
Diseño	40 horas x 17 €/hora	680€
Desarrollo	160 horas x 17 €/hora	2.720€
Documentación	20 horas x 17 €/hora	340€
Licencias de Software	Utilización de software libre y herramientas gratuitas	0€
Infraestructura Tecnológica	Desarrollo realizado en un portátil personal	0€
Total		5.100€

Tabla 8.1: Resumen del Presupuesto Desglosado

Bibliografía

- [1] Bean, J. P. and Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, 55(4):485–540.
- [2] Blackboard (s.f.). *Learning Analytics - Blackboard Academy*. <https://www.blackboard.com/services/consulting/blackboard-academy/learning-analytics>. Consultado el 25 de junio de 2024.
- [3] El Debate (2023). El fracaso escolar generó un sobre coste de más de 4.500 millones de euros en España el curso pasado. https://www.eldebate.com/educacion/20231030/fracaso-escolar-genero-sobrecoste-4500-millones-euros-espana-curso-pasado_150049.html. Consultado el 4 de julio de 2024.
- [4] Git (s.f.). *Git*. <https://git-scm.com>. Consultado el 30 de junio de 2024.
- [5] IntelliBoard (s.f.). *Higher Education Solutions*. <https://intelliboard.net/solutions/higher-education/>. Consultado el 23 de junio de 2024.
- [6] Kocbek, A. and Juric, M. B. (2010). Using advanced business intelligence methods in business process management. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, Ljubljana, Slovenia.
- [7] Kovacic, Z. (2010). Early prediction of student success: Mining students enrolment data. pages 647–665.
- [8] Learning Pool (s.f.). *Learning Pool*. <https://learningpool.com/>. Consultado el 23 de junio de 2024.
- [9] Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development*, 2(4):314–319.
- [10] Marbouti, F., Diefes-Dux, H. A., and Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers Education*, 103:1–15.
- [11] Microsoft (s.f.). *Visual Studio Code*. <https://code.visualstudio.com>. Consultado el 30 de junio de 2024.
- [12] Ministerio de Universidades (s.f.). Metodología de los indicadores universitarios de rendimiento académico. https://www.universidades.gob.es/wp-content/uploads/2023/08/Metodologia_IRA.pdf. Consultado el 30 de junio de 2024.

- [13] Ministerio de Universidades. María Fernández Mellizo (2022). *Análisis del abandono de los estudiantes de grado en las universidades presenciales en España*. https://www.universidades.gob.es/wp-content/uploads/2022/11/EAU_Informe_abandono.pdf. Consultado el 30 de junio de 2024.
- [14] Open Education Analytics (2024). Open education analytics. <https://openeducationanalytics.org>. Consultado el 9 de julio de 2024.
- [15] Perilli, F. (2024). Business intelligence universidades. <https://github.com/Fabrizzioperilli/Business-Intelligence-Universidades.git>. Consultado el 9 de julio de 2024.
- [16] PgAdmin (s.f.). *PgAdmin*. <https://www.pgadmin.org>. Consultado el 30 de junio de 2024.
- [17] Plotly (2023). Scaling your app. <https://dash.plotly.com/dash-enterprise/scale?de-version=5.3>. Consultado el 4 de julio de 2024.
- [18] Plotly (2024). Deploy your dash app. <https://dash.plotly.com/deployment>. Consultado el 4 de julio de 2024.
- [19] Plotly. (s.f.). *Dash user guide*. <https://dash.plotly.com/>. Consultado el 25 de junio de 2024.
- [20] PostgreSQL (s.f.). *PostgreSQL Documentation*. Consultado el 1 de julio de 2024.
- [21] Santi, R. P. and Putra, H. (2018). A systematic literature review of business intelligence technology, contribution and application for higher education. In *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 404–409.
- [22] Saryerwinnie, J. (s.f.). Faker documentation. <https://faker.readthedocs.io/en/master/>. Consultado el 1 de julio de 2024.
- [23] Scikit-learn (s.f.). *Scikit-learn: Machine Learning in Python*. <https://scikit-learn.org/stable/>. Consultado el 30 de junio de 2024.
- [24] SQLAlchemy (s.f.). The python sql toolkit and object relational mapper. <https://www.sqlalchemy.org>. Consultado el 1 de julio de 2024.
- [25] Tinto, V. (1987). *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press.
- [26] ULLaudiovisual. Universidad de La Laguna (2024). Jornadas sobre abandono académico universitario: nuevos retos y desafíos. día 2. <https://www.youtube.com/live/PBRRteHLJNU?feature=shared&t=12000>. Consultado: 1 de julio de 2024.
- [27] Universidad de La Laguna (2024). I jornadas sobre abandono académico universitario: nuevos retos y desafíos. <https://www.ull.es/porta1/agenda/evento/35449/>. Consultado: 1 de julio de 2024.
- [28] Universidad de La Laguna (s.f.). *Portal de Ciencia. Proyecto de investigación PID2020-114849RB-I00*. <https://portalciencia.ull.es/proyectos/51782/detalle>. Consultado el 30 de junio de 2024.

[29] Watershed (s.f.). *Watershed LRS*. <https://www.watershedlrs.com>. Consultado el 22 de junio de 2024.