APLICACIÓN WEB PARA LA PREPARACIÓN Y LIMPIEZA DE DATOS DE ENCUESTAS A EGRESADOS DE LA UNIVERSIDAD POPULAR DEL CESAR SECCIONAL AGUACHICA.

CARLOS ALBERTO MEJIA RODRIGUEZ
FERNANDO GONZALEZ MORA

UNIVERSIDAD POPULAR DEL CESAR
FACULTAD DE INGENIERIAS Y TECNOLOGIAS
ESPECIALIDAD EN INGENIERÍA DE SOFTWARE
LÍNEA DE INVESTIGACIÓN INGENIERÍA DE SOFTWARE
VALLEDUPAR, CESAR
2024

APLICACIÓN WEB PARA LA PREPARACIÓN Y LIMPIEZA DE DATOS DE ENCUESTAS A EGRESADOS DE LA UNIVERSIDAD POPULAR DEL CESAR SECCIONAL AGUACHICA.

CARLOS ALBERTO MEJIA RODRIGUEZ
FERNANDO GONZALEZ MORA

Propuesta de proyecto de grado para obtener el título de Especialista en Ingeniería de Software.

Director
Mag. Luis Manuel Palmera Quintero

UNIVERSIDAD POPULAR DEL CESAR
FACULTAD DE INGENIERÍAS Y TECNOLÓGICAS
ESPECIALIDAD EN INGENIERIA DE SOFTWARE
LÍNEA DE INVESTIGACIÓN INGENIERIA DE SOFTWARE
VALLEDUPAR, CESAR
2024

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

1. DESCRIPCIÓN GENERAL

1.1 TÍTULO DEL PROYECTO

Desarrollo de una aplicación web para la preparación y limpieza de datos de encuestas de Egresados en la Universidad Popular del Cesar Seccional Aguachica.

1.2 DIRECCIÓN DE EJECUCIÓN DEL PROYECTO

CRA 40 VIA AL MAR, Aguachica, Colombia.
egresados.aguachica@unicesar.edu.co

1.3. LAPSO DE EJECUCIÓN DEL PROYECTO

6 meses.

1.4. ORGANISMO Y SECCIÓN RESPONSABLE

Oficina de Egresados de la Universidad Popular del Cesar Seccional Aguachica

1.5. INFORMACIÓN DE CONTACTO DE LOS ESTUDIANTES

Tabla 1
Información de contacto de los estudiantes

Nombre	Apellido	Cédula	Teléfono	Correo
Carlos Alberto	Mejía Rodríguez	1065875988	3184535544	calbertomejia@unicesar.edu.co
Fernando	González Mora	9694608	3184489440	fernando.gonzalez@unicesar.edu.co

Nota: Elaboración propia.

2. DESCRIPCIÓN SITUACIONAL

2.1. IDENTIFICACIÓN DEL PROBLEMA

La Oficina de Seguimiento a Egresados de la Universidad Popular del Cesar Seccional Aguachica tiene como objetivo principal fortalecer la relación entre la institución educativa y sus graduados. Esto implica mantener actualizada la información de los egresados, fomentar su participación en actividades de investigación, apoyar su inserción laboral y ofrecerles oportunidades de educación posgradual. La gestión de datos de los egresados es un componente fundamental de esta oficina, y para recopilar dichos datos, los titulados deben completar diversas encuestas, algunas solicitadas por el Observatorio Laboral para la Educación (OLE) del Ministerio de Educación Nacional (MEN) y otras diseñadas particularmente por la Oficina de Egresados.

La situación actual en el departamento de Egresados respecto a los datos recolectados por las encuestas se caracteriza por dificultades en su preparación y limpieza. La variabilidad en la estructura de las encuestas y la presencia ocasional de datos de baja calidad, caracterizados por valores faltantes, atípicos o erróneos, representan desafíos significativos que afectan la integridad y utilidad de la información recopilada. Esta situación dificulta el aprovechamiento efectivo de los datos. Estos datos son fundamentales para realizar seguimiento y comprender las necesidades y expectativas de los egresados, pero el proceso actual de limpieza es manual, lento y propenso a errores, lo que conduce a conjuntos de datos incompletos y poco confiables.

Si esta situación persiste, la Universidad enfrentará serias dificultades en el análisis y utilización efectiva de los datos de las encuestas. Esto puede desencadenar decisiones poco fundamentadas, una comprensión limitada de las necesidades de los egresados y una disminución en la calidad de los servicios y programas ofrecidos. Además, el proceso manual y lento aumenta la carga de trabajo del personal y retrasa la disponibilidad de los datos para su uso.

Ante este panorama, se propone el desarrollo de una aplicación web que permita automatizar el proceso de preparación y limpieza de datos de encuestas. Para lograr el objetivo se considerarán las primeras fases del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), un enfoque ampliamente reconocido en la minería de datos. Estas fases, que incluyen la comprensión del negocio y la comprensión de los datos, serán cruciales para identificar patrones y entender la naturaleza de los datos recolectados. Luego, se procederá con la preparación de datos, donde se realizará la limpieza, transformación y adecuación de estos. La solución propuesta permitirá la carga de los resultados de las encuestas, para luego seguir con la identificación de columnas numéricas y alfanuméricas, la detección y corrección automatizada de valores, proporcionando opciones eficientes para garantizar la calidad de los datos. Esto facilitaría un análisis más efectivo y la toma de decisiones informadas en la Oficina de Egresados, optimizando el tiempo del personal y mejorando la disponibilidad de los datos para su uso.

En respuesta al escenario planteado, se propone el diseño y desarrollo de una aplicación web con funcionalidades avanzadas para automatizar el proceso de preparación y limpieza de los datos recopilados mediante las encuestas. Para alcanzar este objetivo, se estudiarán y considerarán las primeras fases del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), un enfoque consolidado en la minería de datos. Estas fases, que comprenden la comprensión del negocio y la estructura de los datos, desempeñarán un papel crucial en la identificación de patrones y la comprensión profunda de la naturaleza de los datos recolectados. Posteriormente, se procederá con la etapa de preparación de datos, que abarca actividades críticas como la limpieza, transformación y normalización de estos. La solución propuesta permitirá la importación eficiente de los resultados de las encuestas, seguida de una identificación precisa de las columnas numéricas y alfanuméricas, junto con la detección y corrección automatizada de valores anómalos. Estas características están diseñadas para garantizar la calidad y coherencia de los datos, facilitando así un análisis más profundo y una toma de decisiones informada en la Oficina de Egresados. Además, esta iniciativa se presenta como una alternativa sólida para superar las dificultades actuales en la gestión y análisis de datos, ofreciendo una solución eficaz y eficiente que optimiza el tiempo del personal y mejora la disponibilidad de los datos para su uso.

2.2. JUSTIFICACIÓN DE PROYECTO

La justificación del proyecto se sustenta en diversos aspectos que abordan tanto la necesidad teórica como práctica de desarrollar una solución de software para la preparación y limpieza de datos en la Oficina de Egresados de la Universidad Popular del Cesar Seccional Aguachica. Desde una perspectiva teórica, se busca contribuir al avance académico mediante la aplicación de conceptos de Ciencia de Datos como las matemáticas, estadística y la programación de algoritmos para reemplazar métodos manuales con una solución software especializada que automatice buena parte de las tareas de optimización de los datos. Esto no solo mejorará la eficiencia del proceso, sino que también resaltará las ventajas de utilizar software y ciencia de datos en el ámbito de la identificación y corrección de inconsistencias en la información almacenada. Asimismo, al proporcionar una herramienta que pueda generalizar resultados a principios más amplios, se espera demostrar cómo mejorar la calidad y utilidad de los datos recolectados en la Oficina de Egresado puede tener implicaciones significativas en otros ámbitos de investigación y aplicación.

Desde un punto de vista práctico, el proyecto responde a la necesidad de abordar problemas reales que impactan la calidad y confiabilidad de los datos almacenados, específicamente los recolectados por la Oficina de Egresados a través de encuestas. Al proponer una solución que automatice eficientemente ciertas etapas del proceso de limpieza de datos, no solo se busca solventar las dificultades actuales en la preparación y optimización de los datos recolectados, sino también introducir estrategias basadas en técnicas efectivas que mejoren la depuración de la información y, por ende, reduzcan el tiempo usualmente requerido para tal fin. La implementación de esta herramienta práctica podría impactar significativamente en la toma de decisiones informadas y en la calidad de los servicios y programas ofrecidos a los egresados de la universidad, dado que la calidad de dichos servicios y programas depende en gran medida de la calidad de los datos recolectados y procesados.

Desde una perspectiva metodológica, el proyecto se enriquece mediante la combinación de metodologías de investigación, destacando un enfoque cualitativo principalmente, así como técnicas de minería de datos como el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), y un enfoque ágil de desarrollo de software, específicamente Scrum, para guiar el desarrollo del software destinado a la preparación y limpieza de datos en la Oficina de

Egresados. La selección de un enfoque cuantitativo se justifica por la necesidad de medir de manera numérica los errores presentes en las columnas del conjunto de datos recolectados por la Oficina de Egresados, lo cual facilitará una evaluación precisa de la calidad de los datos y la identificación de posibles mejoras en el proceso de limpieza. Además, se opta por el modelo KDD debido a que sus primeras fases permiten realizar la limpieza, transformación y preparación de los datos de manera sistemática y eficiente, asegurando así que los datos estén listos para su análisis y aprovechamiento posterior. La adopción de un enfoque ágil de desarrollo de software como Scrum garantizará una gestión flexible y adaptativa del proyecto, permitiendo una respuesta rápida a los cambios y mejoras durante todo el proceso de desarrollo. La combinación de estas metodologías y enfoques técnicos no solo promoverá el avance académico en el campo de la gestión de datos, sino que también se espera que la solución informática resultante tenga un impacto considerable en la mejora de la calidad de los datos, los cuales son insumos vitales para los servicios ofrecidos a los egresados de la Universidad Popular del Cesar Seccional Aguachica.

2.3. OBJETIVOS DEL PROYECTO

2.3.1 Objetivo General

Desarrollo de una aplicación web para la preparación y limpieza de datos de encuestas a egresados de la universidad popular del cesar seccional Aguachica

2.3.2 Objetivos específicos

- Analizar los requisitos específicos de la aplicación, considerando las necesidades de la Oficina de Egresados y los estándares de limpieza de datos.
- Diseñar el módulo de limpieza de datos, que incluya funcionalidades para importar los datos, identificar y corregir inconsistencias, duplicados y datos faltantes.
- Crear el módulo de generación de reportes, que posibilite la visualización y exportación de los datos limpios.
- Implementar la aplicación web en un entorno de producción para asegurar la eficiente preparación y limpieza de los datos de las encuestas de egresados.

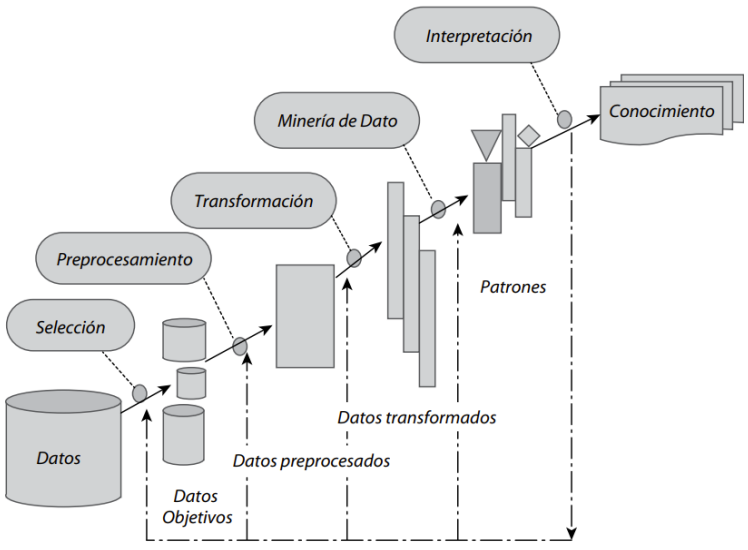
2.4 MARCO METODOLÓGICO

El proyecto se basa en la combinación de métodos y técnicas para adquirir, analizar e interpretar resultados. Se emplea un enfoque cuantitativo que posibilita evaluar el nivel inicial de inconsistencias en los datos obtenidos de las encuestas a egresados y medir el impacto de la optimización. El proceso de limpieza de datos se basa en métodos numéricos y estadísticos el enfoque cuantitativo permitirá garantizar la precisión y fiabilidad de los resultados obtenidos.

Los administradores de proyectos de minería de datos se benefician al usar un modelo estándar, como el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD, siglas en inglés de Knowledge Discovery in Databases), ya que reducen costos y tiempos, facilitan la transferencia de conocimientos y promueven la reutilización de las mejores prácticas [1].

El proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), representado en la Figura 1, es un proceso interactivo e iterativo que implica múltiples pasos. Se puede resumir en las siguientes etapas: selección, preprocesamiento/limpieza, transformación/reducción, minería de datos y finalmente, interpretación/evaluación.

Figura 1
Etapas del proceso KDD.



Fuente: propia de los autores.

Scrum es una metodología ágil que resulta muy útil para la limpieza y preparación de datos, fases iniciales del proceso KDD. Según [2], Scrum organiza el trabajo en sprints iterativos, fomentando la colaboración y la entrega continua de valor. Con roles claramente definidos (Product Owner, Scrum Master y equipo de desarrollo), eventos regulares (reuniones diarias, revisiones y retrospectivas) y artefactos (backlog de producto y backlog de sprint), todo ello proporciona una estructura eficiente para guiar el desarrollo de proyectos.

Para el desarrollo de la plataforma se adopta una arquitectura que separa claramente las capas de backend y frontend, garantizando la independencia de la lógica empresarial y la presentación de la interfaz de usuario. El backend se construye utilizando Django y Python, un robusto marco de trabajo que facilita el desarrollo rápido y seguro de la aplicación web, permitiendo la gestión eficiente de la base de datos, el manejo de rutas y vistas, así como la seguridad. En cuanto al frontend, se utilizan HTML, CSS y JavaScript, junto con el framework Bootstrap para diseñar interfaces modernas y responsivas. Complementariamente, se incorporan bibliotecas como Matplotlib, Seaborn y Pandas para la generación de gráficos y el análisis de datos, mientras que Openpyxl y ReportLab permiten la creación y manipulación de reportes en Excel y PDF. Por último, se emplea SQLite como sistema de gestión de bases de datos para almacenar y organizar la información, facilitando la integración con el resto de los componentes de la aplicación.

La aplicación web debe contar con la capacidad de importar archivos de distintos formatos o fuentes, tales como CSV, XLSX o SQL, los cuales serán convertidos en un dataframe utilizando Python (lenguaje de programación) para dar inicio al procesamiento de los datos, con el objetivo de garantizar la máxima calidad de estos. Posteriormente, se habilitará la funcionalidad para exportar los datos en cualquiera de los formatos deseados.

3. DESARROLLO CIENTÍFICO TECNOLÓGICO

3.1 MARCO REFERENCIAL

El marco referencial del proyecto enfoca en definir los conceptos, metodologías, normas y técnicas que respaldan el desarrollo de una aplicación web para la preparación y limpieza de datos de encuestas a egresados de la Universidad Popular del Cesar Seccional Aguachica. En este contexto, se parte de un análisis de los antecedentes relacionados con la gestión de datos y la calidad de estos en entornos educativos, seguido de una revisión teórica sobre los procesos de limpieza de datos, incluyendo conceptos clave como la minería de datos, el Descubrimiento de Conocimiento en Bases de Datos (KDD), y las metodologías ágiles como Scrum. Además, se exploran las herramientas tecnológicas y arquitecturas recomendadas para la implementación de aplicaciones web, así como las buenas prácticas en el manejo de grandes volúmenes de información. Por último, se abordan las regulaciones y estándares aplicables para garantizar la integridad, seguridad y calidad de los datos

3.1.1 Antecedentes

En cuanto a la búsqueda de investigaciones realizadas con relación a aplicaciones web para la preparación y limpieza de datos en ambientes educativos, se exponen los siguientes resultados a nivel internacional, nacional y local.

3.1.1.1 Antecedentes Internacionales

En el apartado internacional, el estudio de [3], se centra en la aplicación de un modelo de analítica de texto para detectar y clasificar automáticamente las tendencias de investigación en e-learning, con un enfoque específico en la Universidad Oberta de Catalunya. Este estudio aplicó técnicas de analítica de texto y minería de texto para detectar y clasificar tendencias de investigación en e-learning. Utilizaron un modelo de clasificación basado en reglas semánticas de categorización profunda. Se evaluó la precisión del modelo en diferentes categorías y niveles de análisis. Los resultados mostraron la viabilidad de la minería de texto para identificar tendencias de investigación en este campo, aunque se destacó la necesidad de considerar

matices contextuales para mejorar la precisión de la clasificación automática.

En investigaciones internacionales, el artículo de [4], titulado "Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review," se centra en la evaluación de profesores en la educación superior y cómo se pueden utilizar técnicas de minería de datos para predecir esta evaluación. El estudio se basa en una revisión sistemática de la literatura y tiene como objetivo identificar estudios que integren aplicaciones predictivas mediante la minería de datos educativos, centrándose en la evaluación de profesores y el rendimiento de los estudiantes y se analizan distintas técnicas, como la lógica difusa, la clasificación y otras herramientas de análisis de datos.

3.1.1.2 Antecedentes Nacionales

En plano nacional se resalta el trabajo de grado de [5], el cual tenía como objetivo principal aplicar el proceso de Descubrimiento de Conocimiento en Datos (KDD) a datos educativos, específicamente a información académica y no académica de graduados de programas de posgrado (maestrías, especializaciones y doctorados) en la Universidad Nacional de Colombia. Se enfocó en la etapa de minería de datos con dos objetivos principales: encontrar relaciones entre los datos de los graduados y el tiempo que les tomó completar sus estudios, y desarrollar un modelo predictivo para estudiantes que ingresen a estos programas en el futuro. Además, se buscó proporcionar una visualización de los resultados para que los programas de posgrado puedan tomar decisiones basadas en la situación de cada uno de ellos al planificar mejoras.

Dentro de las investigaciones a nivel nacional, se resalta el proyecto desarrollado por [6] como parte de su Maestría en Software Libre, que se centra en mejorar la predicción de la deserción académica de estudiantes de pregrado en la Universidad Autónoma de Bucaramanga mediante el empleo de técnicas de minería de datos. Este estudio hace uso de herramientas de código abierto, como Weka, y el algoritmo de clasificación J48 con el objetivo de optimizar el modelo de estimación de riesgo de deserción en la institución educativa.

3.1.1.3 Antecedentes Locales

Un estudio relevante a nivel local es el de [7], quienes analizaron la relación entre la implementación de un programa de orientación vocacional en las instituciones de educación media del Departamento del Cesar y la deserción estudiantil en la Universidad Popular del Cesar (UPC). Utilizando un diseño correlacional no experimental, los autores encontraron una asociación moderada entre las variables, destacando la importancia de la orientación vocacional para reducir la tasa de deserción en la universidad. Los resultados sugieren que una mejor orientación en la elección de carrera podría contribuir a la retención de estudiantes en la UPC, y servir de base para estrategias preventivas futuras en la institución educativa.

A nivel regional se trae a mención la investigación de [8], la cual trata sobre el desarrollo de un software que utiliza técnicas de minería de datos para predecir la deserción estudiantil en el programa de Ingeniería de Sistemas de la Universidad Francisco de Paula Santander Seccional Ocaña. Se parte de una revisión de la literatura sobre deserción estudiantil y estrategias para la predicción del rendimiento académico. El objetivo principal es crear una herramienta que automatice el proceso de limpieza de datos, construya un modelo de predicción y plantee estrategias preventivas para estudiantes en riesgo de deserción. La justificación radica en la necesidad de abordar la deserción estudiantil y mejorar la toma de decisiones en la institución.

3.1.2 Marco Teórico

En este apartado se describen los principales términos que guiarán la construcción de la aplicación, como es la preparación, limpieza y minería de datos. También es clave explicar los modelos y metodologías a utilizar, como el proceso KDD (Knowledge Discovery in Databases) y la metodología Scrum.

3.1.2.1 Preparación y limpieza de datos

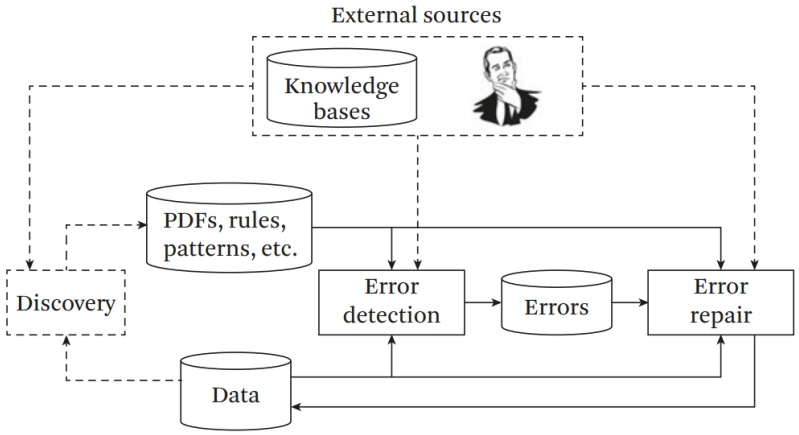
La limpieza o reparación de datos, según [9], se define como un problema crucial en muchas tareas relacionadas con bases de datos. Consiste en hacer que una base de datos sea consistente con respecto a un conjunto de restricciones específicas. Para [10], la preparación

de datos es el proceso que implica organizar, transformar y estructurar los datos brutos para que sean consistentes, completos y adecuados para su análisis posterior. Esto incluye la eliminación de errores, datos faltantes y registros no deseados, así como la reorganización de columnas y la normalización de formatos.

El flujo de limpieza de datos, ilustrado en la Figura 2, comienza con la obtención y análisis de la información para entender su estructura. Luego se identifican errores como datos duplicados o atípicos, y finalmente se aplican correcciones para obtener un conjunto de datos limpio y listo para su uso. Durante todo el proceso, se puede recurrir a expertos o fuentes externas para asegurar la calidad de las correcciones [11].

Figura 2

Flujo típico de limpieza de datos



Nota: Tomado de [11]

La Tabla 1 presenta un resumen de los principales componentes y técnicas utilizadas en el proceso de limpieza de datos, abarcando desde la detección y reparación de errores hasta la utilización de métodos avanzados y la participación humana en la validación de las correcciones.

Tabla 2

Componentes y técnicas utilizadas en la limpieza de datos

Categoría	Descripción
Detección de errores	Detección de valores atípicos, identificación de duplicados y detección basada en reglas de calidad de datos.
Reparación de errores	Resolución de violaciones de integridad, actualización de valores y eliminación de inconsistencias.
Técnicas utilizadas	Enfoques cuantitativos (métodos estadísticos), enfoques cualitativos (patrones y reglas) y transformación de datos para estandarización.
Métodos avanzados	Limpieza guiada por Machine Learning (ML), integración de aprendizaje automático en el proceso de detección y reparación.
Participación humana	Consulta de expertos en casos de alta incertidumbre y verificación manual de metadatos y correcciones propuestas.

Nota. Tabla elaborada a partir de [11].

Para [12], la limpieza de datos se organiza en cuatro dominios principales que son integridad, consistencia, completitud y exactitud, cada uno abarcando tareas específicas como se evidencia en la Tabla 2.

Tabla 3

Lista de tareas para la limpieza de datos

Dominio	Tarea
Integridad de Datos	Crear un diccionario de datos, asegurar que las variables estén correctamente formateadas e importadas; definir la pregunta de investigación; aplicar criterios de inclusión/exclusión; evaluar la necesidad de transformar variables o crear nuevas; y verificar si el conjunto de datos es representativo y apropiado para el estudio.
Consistencia de Datos	Identificar valores incompatibles o inconsistentes entre variables relacionadas y corregirlos.
Completitud de Datos	Identificar y cuantificar datos faltantes; analizar patrones de ausencia y determinar cómo se manejarán en el análisis del estudio.
Exactitud de Datos	Evaluar variables continuas para detectar valores atípicos mediante pruebas estadísticas y visualizaciones; manejar outliers de manera adecuada y realizar un análisis de sensibilidad para verificar el impacto de su eliminación en los resultados.

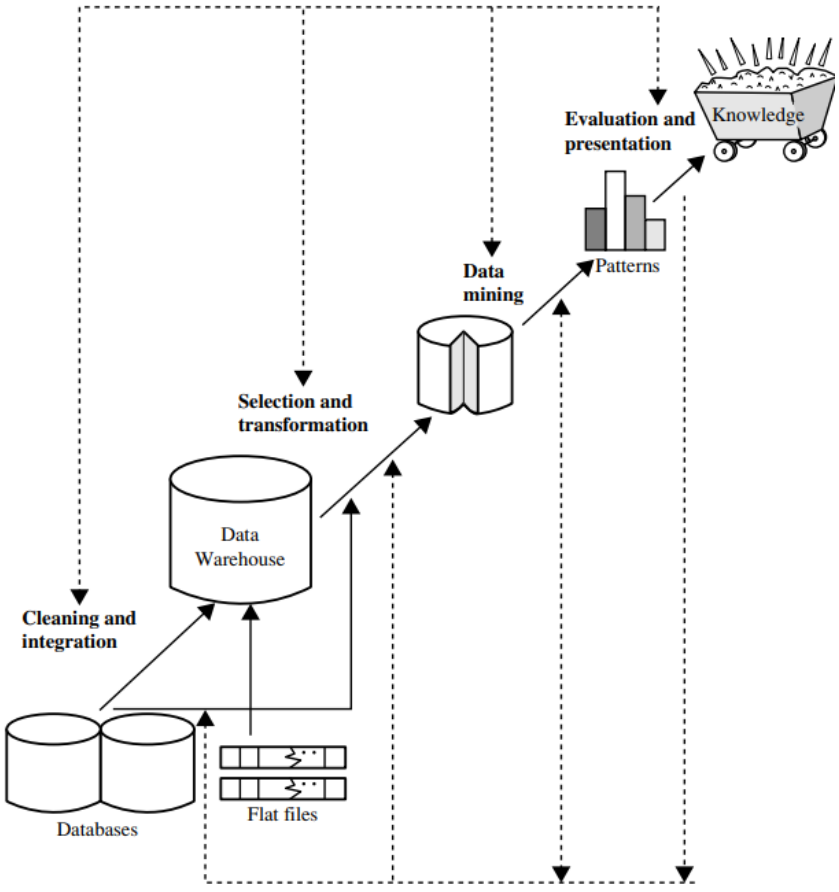
Nota: Tabla elaborada a partir de [12]

3.1.2.2 Minería de datos y KDD

[13] definen la minería de datos como el proceso de descubrir patrones en grandes volúmenes de datos de forma automática o semiautomática. Estos patrones deben ser significativos y proporcionar algún tipo de ventaja, normalmente económica. La minería de datos se centra en encontrar relaciones que permitan realizar predicciones y comprender la estructura de los datos almacenados.

La minería de datos es una parte esencial del proceso KDD, que abarca desde la selección de datos hasta su análisis e interpretación. Mientras la minería se centra en descubrir patrones significativos, el KDD va más allá al aplicar estos hallazgos como conocimiento útil para la toma de decisiones [14]. La Figura 3 ilustra esta integración.

Figura 3
Minería de datos en el proceso de descubrimiento de conocimiento.



Nota: Tomado de [14]

Las fases del KDD (Knowledge Discovery in Databases), según el [15], son las siguientes:

- Selección de datos: Extraer datos relevantes para el análisis desde la base de datos.
- Preprocesamiento de datos: Limpieza y preparación de los datos para asegurar calidad y consistencia.
- Transformación de datos: Convertir los datos a un formato apropiado para la minería.
- Selección de la tarea de minería de datos: Determinar el tipo de análisis que se realizará, como clasificación, regresión, agrupamiento o resumen.
- Selección del algoritmo de minería de datos: Elegir el método específico que se utilizará para buscar patrones.
- Aplicación del algoritmo de minería de datos: Ejecutar el algoritmo elegido para extraer patrones y relaciones en los datos.
- Evaluación e interpretación: Revisar y analizar los patrones descubiertos para determinar la relevancia y aplicabilidad de estos.
- Despliegue: Utilizar el conocimiento obtenido en la práctica, integrarlo en sistemas existentes, o documentarlo para una futura aplicación.

3.1.2.3 Metodología ágil Scrum

Scrum es un marco de trabajo para gestionar productos, proyectos y servicios complejos, que se basa en un desarrollo incremental y sostenido. Es un framework que establece roles, artefactos y actividades específicas dentro de un proyecto, con el propósito de crear un flujo de comunicación efectivo que cubra todas las necesidades del equipo. La forma en que se comunica, a quién y cuándo se hace, juega un papel crucial en el éxito del proyecto, permitiendo que cada miembro del equipo ejerza su rol de manera efectiva y cumpla con sus compromisos profesionales y de equipo [16].

[17] define Scrum como un marco de trabajo que organiza la gestión de proyectos mediante roles claramente definidos, facilitando la colaboración en el equipo. La Tabla 3 muestra estos roles y sus responsabilidades.

Tabla 4

Roles y Responsabilidades en Scrum

Rol	Descripción	Responsabilidades
Product Owner	Persona encargada de definir la visión del producto y gestionar el backlog.	Establecer prioridades en el backlog. Planificar la entrega de características en cada sprint. Asegurar que el equipo aporte valor a los stakeholders.
Scrum Master	Facilitador que se asegura de que Scrum se utilice correctamente y que el equipo funcione eficientemente.	Guiar al equipo y al Product Owner en el uso de Scrum. Eliminar impedimentos que afecten el progreso. Promover mejoras continuas en el proceso de Scrum.
Equipo de Desarrollo	Grupo responsable de ejecutar y entregar el producto final, compuesto por miembros multifuncionales.	Estimar tareas y colaborar con el Product Owner para planificar los sprints. Ejecutar las tareas asignadas y desarrollar el producto. Participar en revisiones y retrospectivas de sprint.

Fuente: Tabla elaborada a partir de [17]

En el marco de Scrum, los artefactos se utilizan para proporcionar transparencia y control en la ejecución de los proyectos, ayudando a todos los miembros del equipo a tener una visión clara del progreso y del estado del trabajo [18].

Tabla 5

Artefactos de Scrum

Artefacto	Descripción
Product Backlog	Lista priorizada de tareas del producto, gestionada por el Product Owner.
Sprint Backlog	Tareas seleccionadas para completar en un sprint.
Incremento	Trabajo completado que está listo para entrega.
Definition of Done	Criterios que definen cuándo una tarea está realmente terminada.

Fuente: Tabla elaborada a partir de [18].

3.1.2.4 Fundamentos de Machine Learning

El aprendizaje automático, también conocido como machine learning, es una rama de la inteligencia artificial que se enfoca en crear sistemas capaces de aprender y mejorar a partir de los datos. Estos sistemas utilizan un proceso de entrenamiento para desarrollar modelos predictivos basados en experiencias. Según [19], hay dos tipos principales de algoritmos: supervisados y no supervisados, que se seleccionan dependiendo del tipo de resultados que se quieran obtener.
Los algoritmos supervisados son especialmente útiles cuando se cuenta con pocas instancias etiquetadas para aprender y una gran cantidad de datos no etiquetados. En estos escenarios, los algoritmos supervisados, como los utilizados en tareas de clasificación y regresión, se convierten en la mejor opción para entrenar modelos y obtener resultados precisos [20].

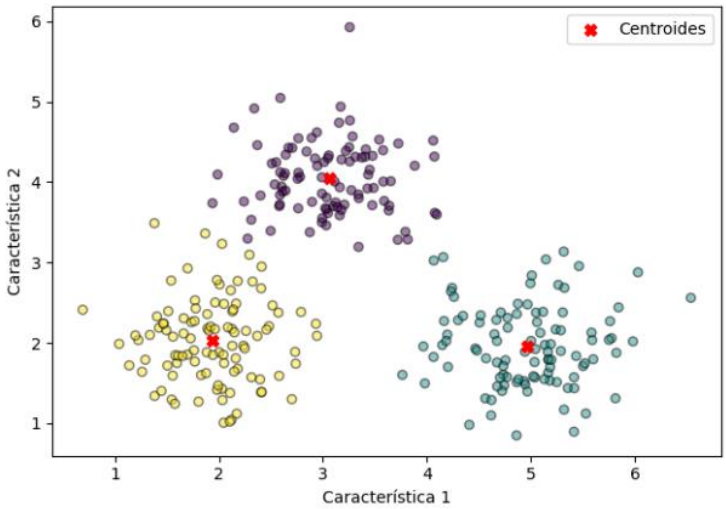
Los algoritmos no supervisados se caracterizan por no tener una meta específica durante el aprendizaje. En lugar de seguir una función objetivo, se enfocan en encontrar patrones y relaciones dentro de los datos. Se utilizan en una amplia variedad de aplicaciones prácticas, como optimizar la disposición de productos en estanterías o detectar fallos mecánicos relacionados entre sí [21].

El clustering es un método de aprendizaje no supervisado que se utiliza para agrupar elementos según sus similitudes, sin que se conozcan previamente los grupos a los que pertenecen. A través de este proceso, las instancias se organizan en distintos conjuntos basados en sus características comunes [22].

El algoritmo K-Means se utiliza para agrupar datos no etiquetados y descubrir patrones ocultos. Su objetivo es dividir los datos en grupos con características similares, asignándolos a diferentes clústeres. Para lograrlo, identifica K puntos centrales, llamados centroides, que representan el centro de cada grupo. Posteriormente, asigna cada dato al clúster más cercano. Este método es especialmente útil cuando se trabaja con datos sin clasificar y se conoce de antemano la cantidad de grupos deseados [23]. Un ejemplo de clustering con K-Means se muestra en la Figura 4, donde se puede observar cómo los datos se agrupan en distintos clústeres según sus características.

Figura 4

Clustering con k-means



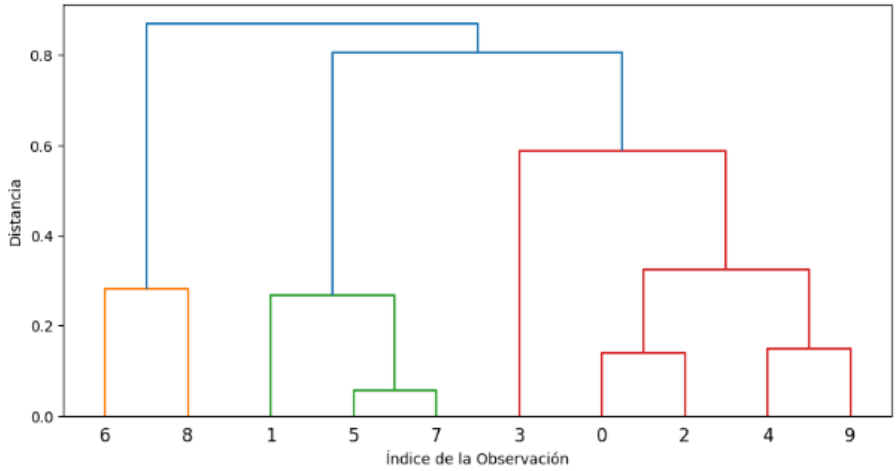
Nota: Elaboración propia

El algoritmo K-Means se aplica en muchos campos prácticos. En el entorno empresarial, se emplea para segmentar mercados, agrupando a los clientes con características similares para personalizar las estrategias según cada segmento. También se utiliza en la clasificación de libros, películas y otros documentos, así como en la detección de fraudes y actividades sospechosas, analizando datos para encontrar patrones y similitudes que ayuden a entender mejor el comportamiento de los clientes o usuarios. [24].

El clustering jerárquico es una técnica de agrupamiento que utiliza conceptos de teoría de grafos y aprendizaje no supervisado para organizar elementos similares de manera jerárquica. Este método ayuda a descubrir y visualizar la estructura jerárquica presente en los datos, mostrando cómo se relacionan entre sí dentro del conjunto analizado. [25]. El clustering jerárquico inicia formando un clúster por cada instancia en el conjunto de datos, de manera que cada clúster contiene una única instancia al principio. Luego, de forma iterativa, se van identificando los dos clústeres más cercanos (según medidas como la distancia euclidiana) y se fusionan en un nuevo clúster. Este proceso continúa hasta que todos se agrupan en un solo clúster final. El resultado es un dendrograma que ilustra cómo se organizan jerárquicamente las instancias [26]. Se presenta un ejemplo en la Figura 5.

Figura 5

Dedograma clustering jerárquico



Nota: elaboración propia.

La selección de características y la reducción de dimensionalidad son aspectos clave del aprendizaje automático, ya que mejoran el rendimiento de los modelos en conjuntos de datos con muchas variables. Estas prácticas no solo optimizan la eficiencia de los algoritmos supervisados, sino que también juegan un papel fundamental en los algoritmos no supervisados, ya que permiten identificar características relevantes y gestionar la dimensionalidad para obtener resultados de mayor calidad [27].

3.2 RESULTADOS Y ANÁLISIS DE RESULTADOS

El desarrollo de la aplicación web para la limpieza y preparación de datos sigue las fases del ciclo de vida del software, alineándose con los objetivos del proyecto y organizando las actividades en sprints cortos bajo la metodología Scrum. A continuación, se presentan los resultados de cada etapa del ciclo, describiendo las actividades y avances en cada fase.

3.2.1 Planificación del Proyecto

Según [28], la planificación es la primera fase en el desarrollo de software y su propósito es definir objetivos, alcance y limitaciones. Esta fase proporciona una base para que los stakeholders tomen decisiones informadas y gestionen el proyecto de manera eficiente,

identificando problemas y mitigando riesgos desde el inicio. Incluye una visión general del proceso, organización, entregables, y manejo de recursos, que guía el desarrollo inicial.

3.2.1.1 Identificación del proceso de encuestas a egresados de la Seccional

Se realizó un sondeo preliminar al líder de la Oficina de Egresados de la Universidad Popular del Cesar Seccional Aguachica para evaluar el proceso actual de recolección y uso de datos. La oficina lleva a cabo entre una y cinco encuestas anuales con el propósito de actualizar la base de datos y mejorar los servicios que brindan a los egresados. Dichas encuestas buscan clasificar a los graduados según sus intereses laborales y académicos, para facilitar el desarrollo de estrategias de intervención más efectivas y alineadas con necesidades específicas. La tabla 6 resume los principales hallazgos y aspectos esenciales del proceso de encuestar a los egresados de la Seccional.

Tabla 6

Identificación del Proceso de Encuestas a egresados de la Seccional

Aspecto	Descripción
Frecuencia de Encuestas	Se realizan entre 1 a 5 encuestas al año.
Responsable	Oficina de Egresados de la Seccional.
Propósito de las Encuestas	Actualizar la base de datos de egresados, seguimiento y retroalimentación, mejora de servicios, networking.
Utilización de los Datos	Categorización de egresados según preferencias laborales y académicas.
Acceso a los Datos	Incluyen información personal, académica, posgrados y laborales, accesible para uso institucional.
Métodos de Recolección	Encuestas en línea, entrevistas telefónicas, visitas en persona y otros medios electrónicos.

Nota: Elaboración propia

3.2.1.2 Objetivos del desarrollo

El proyecto tiene como objetivo desarrollar un sistema web que permita importar los resultados de las encuestas, así como automatizar el almacenamiento, limpieza, preparación, análisis y presentación de estos datos. El sistema incluirá herramientas para la visualización y análisis de datos, permitiendo generar estadísticas y reportes detallados. Además, proporcionará funcionalidades para la limpieza y preparación de datos, asegurando su calidad antes del análisis. También permitirá un acceso controlado a los datos para distintos usuarios y garantizará la seguridad y privacidad de la información.

Tabla 7

Funcionalidades propuestas para el Sistema

Característica	Descripción
Importación de Resultados	Permite cargar los resultados de las encuestas desde distintas fuentes y formatos, facilitando la integración de la información.
Limpieza y Preparación de Datos	Automatiza la limpieza y preparación para garantizar datos consistentes y de calidad antes de su análisis.
Análisis de Datos	Herramientas para clasificar y visualizar datos, generando reportes y estadísticas sobre los egresados.
Acceso y Gestión de Datos	Control de acceso para distintos usuarios (personal administrativo y actores relevantes).
Seguridad y Privacidad	Protección de datos personales y cumplimiento de normativas legales y éticas en la gestión de información.

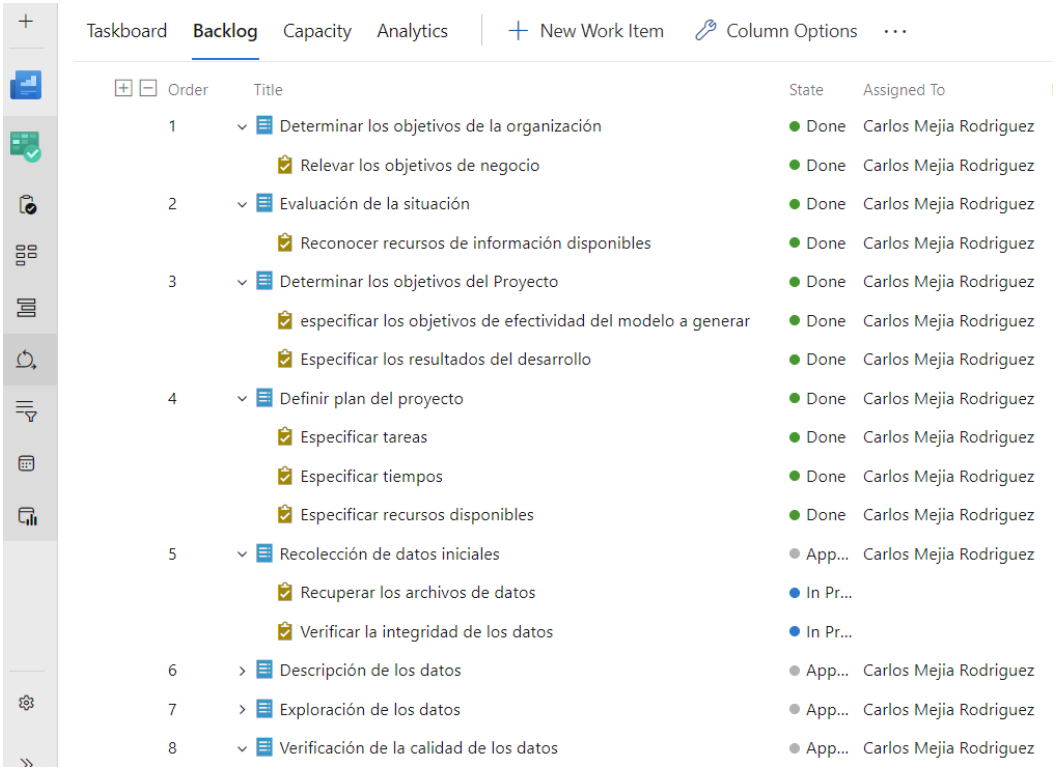
Nota: Elaboración propia

3.2.1.3 Definición del plan del proyecto

En la industria del software, se recomienda utilizar enfoques ágiles como Scrum, ya que permiten organizar el trabajo en tareas generales y descomponerlas en actividades específicas, facilitando la adaptación continua [29]. Por ello, se combinan Scrum con las primeras fases del proceso KDD (selección, limpieza y preparación de datos). Estas etapas se gestionan con Scrum usando herramientas como Azure DevOps, donde se definen Epics, Product Backlog y Tasks. La Figura 6 muestra la planificación de tareas en el Sprint Backlog.

Figura 6

Sprint backlog. Planificación de tareas.



Nota: Planificación hecha en Azure DevOps.

3.2.1.4 Recursos materiales requeridos para el proyecto

El desarrollo del sistema web para la limpieza y preparación de datos requiere diversos recursos materiales para garantizar un entorno de trabajo eficiente. La Tabla 8 detalla los costos y elementos necesarios.

Tabla 8

Presupuesto de Recursos Materiales para el Proyecto

Elementos	Descripción / Servicios Adicionales Requeridos	Cantidad	Valor Unitario	Valor Total
Computador Portátil	Laptop con capacidad de procesamiento adecuada para ejecutar herramientas de análisis de datos y desarrollo web.	1	$2.650.900	$2.650.900
Impresiones	Copias de documentos relacionados con el desarrollo y diseño del proyecto.	100	$100	$10.000
Rodamiento	Movilización de personal y materiales para reuniones y validación de requerimientos.	-	$100.000	$100.000
Internet	Conexión a Internet estable para el desarrollo y acceso a servicios en la nube.	6 meses	$50.000	$300.000
Total				$3.060.900

Nota: Elaboración propia

3.2.1.5 Requisitos funcionales

[30] autores definen los requisitos funcionales como las características o capacidades específicas que una entidad (por ejemplo, un sistema) debe tener para cumplir con las necesidades de los usuarios y del negocio. Son importantes porque describen las acciones que el sistema debe realizar y garantizan que las funcionalidades desarrolladas se alineen con los objetivos de la organización.

Las tablas siguientes describen claramente los requisitos funcionales necesarios para el correcto funcionamiento del sistema web de limpieza y preparación de datos.

La Tabla 9 muestra el requerimiento funcional que detalla cómo debe realizarse la recopilación de datos en el sistema web, incluyendo la admisión de archivos CSV o XLSX y la confirmación de una carga exitosa o mensaje de error.

Tabla 9

Requerimiento Funcional 01 Recopilación de Datos

Código:	RF01
Nombre:	Recopilación de Datos
Propósito:	La aplicación debe permitir la entrada de datos provenientes de encuestas a egresados.
Descripción:	El sistema tendrá la opción de cargar o subir datos de las encuestas de egresados, tipo de fichero admitido CSV o XLSX (Excel).
Entradas:	Archivo CSV o XLSX con registros de encuestas a egresados (datos en formato tabular, columnas separadas por comas y filas por saltos de línea).
Salidas:	Confirmación de carga exitosa o mensaje de error durante la carga de datos.
Prioridad:	Alta.

Nota: Elaboración propia

La Tabla 10 detalla las funcionalidades del sistema para procesar y mejorar la calidad de los datos, garantizando que se corrijan las inconsistencias antes del análisis.

Tabla 10

Requerimiento Funcional 02 Procesamiento de Calidad de Datos

Código:	RF02
Nombre:	Procesamiento de Calidad de Datos
Propósito:	La aplicación debe brindar una interfaz que permita identificar y corregir valores faltantes, atípicos o erróneos en las encuestas de egresados.
Descripción:	La aplicación debe automatizar la identificación y corrección de inconsistencias en los datos de encuestas para garantizar calidad y coherencia, mediante acciones como permutar valores por moda o media, o eliminar registros afectados
Entradas:	Instrucciones de procesamiento de datos.
Salidas:	Confirmación de procesamiento y creación de un nuevo dataset de calidad
Prioridad:	Alta.

Nota: Elaboración propia

La Tabla 11 describe el Requerimiento Funcional 03, enfocado en la selección de datos relevantes para el proceso KDD, facilitando la identificación de las variables más influyentes.

Tabla 11

Requerimiento Funcional 03 Selección de Datos Relevantes para KDD

Código:	RF03
Nombre:	Selección de Datos Relevantes para KDD
Propósito:	Implementar un mecanismo automático que utilice técnicas de análisis de relevancia para sugerir las variables más influyentes en la toma de decisiones y descubrimiento de patrones.
Descripción:	La aplicación debe proporcionar una funcionalidad que identifique y muestre las columnas o variables más relevantes en las bases de datos de encuestas de egresados Procesos de descubrimiento de conocimiento (KDD).
Entradas:	Configuración de parámetros que guiarán el sistema en la identificación de variables relevantes. Esto puede incluir umbrales de importancia, métodos específicos de análisis de relevancia, entre otros.
Salidas:	Un conjunto de variables identificadas como las más relevantes para el proceso KDD, ya sea mediante un análisis automático de relevancia o seleccionadas manualmente por el usuario.
Prioridad:	Alta.

Nota: Elaboración propia

La Tabla 12 presenta el Requerimiento Funcional 04, que se centra en la transformación de datos para que estén en el formato adecuado para la minería de datos y análisis avanzado.

Tabla 12

Requerimiento Funcional 04 Transformación de Datos para Minería

Código:	RF04
Nombre:	Transformación de Datos para Minería
Propósito:	La aplicación debe incluir funcionalidades para transformar los datos de las encuestas de egresados en formas adecuadas para la minería de datos.
Descripción:	Proporcionar herramientas de normalización, estandarización y combinación de variables para asegurar la consistencia y comparabilidad de datos provenientes de las encuestas.
Entradas:	Conjunto de datos de las encuestas de egresados después de pasar por las fases iniciales de limpieza y preprocesamiento. Configuración de parámetros que guiarán Proceso de conversión de datos.

Salidas:	Conjunto de datos resultante después de aplicar las operaciones de transformación. Estos datos estarán en una forma adecuada para las operaciones de minería de datos, con valores limpios, normalizados y preparados para el análisis.
Prioridad:	Alta.

Nota: Elaboración propia

La Tabla 13 detalla el Requerimiento Funcional 05, que proporciona funcionalidades de visualización avanzada para mostrar de manera clara y comprensible los resultados del análisis y minería de datos.

Tabla 13

Requerimiento Funcional 05 Visualización de Resultados

Código:	RF05
Nombre:	Visualización de Resultados
Propósito:	La aplicación debe ofrecer funcionalidades de visualización avanzada para presentar de manera clara y comprensible los resultados del proceso KDD
Descripción:	El sistema deberá Implementar gráficos interactivos, resúmenes visuales y cuadros de mando que destaquen patrones, tendencias y hallazgos relevantes.
Entradas:	Archivo que contienen la información recopilada a en encuestas realizadas a egresados; ya procesado y parámetros específicos del proceso KDD.
Salidas:	Conjunto de datos procesados y analizados mediante técnicas de minería de datos. Gráficos, cuadros de mando y resúmenes visuales que representan de manera clara y comprensible los resultados del proceso KDD.
Prioridad:	Alta.

Nota: Elaboración propia

La Tabla 14 muestra el Requerimiento Funcional 06, que detalla las características para la exportación de los resultados de las primeras fases del proceso KDD en diversos formatos.

Tabla 14

Requerimiento Funcional 06 Exportación de Resultados

Código:	RF06
Nombre:	Exportación de Resultados
Propósito:	Facilitar la exportación de los resultados del proceso KDD en diversos formatos (CSV, Excel, PDF) para su posterior análisis o presentación.
Descripción:	La aplicación tendrá botones claros en su interfaz principal para exportar resultados en formatos como CSV, Excel y PDF. Los usuarios podrán personalizar la exportación, aplicar filtros y recibir notificaciones visuales sobre el progreso. Esta funcionalidad asegura una exportación sencilla y rápida de los resultados del proceso KDD.
Entradas:	Elección del formato y filtros aplicados a los datos de exportación.
Salidas:	Archivo en el formato seleccionado (CSV, Excel, PDF) que contiene los resultados del proceso KDD, con la información solicitada y configurada por el usuario.
Prioridad:	Alta.

Nota: Elaboración propia

3.2.1.6 Requerimientos No Funcionales

Según [30], los requisitos no funcionales se enfocan en la calidad y las restricciones del sistema, como la seguridad y el rendimiento.

La Tabla 15 presenta el Requerimiento No Funcional 01, que se centra en la accesibilidad del sistema, garantizando que el diseño de la aplicación se adapte a diferentes tamaños de pantalla y dispositivos.

Tabla 15

Requerimiento No funcional 01 Accesibilidad

Código:	RNF01
Nombre:	Accesibilidad
Descripción:	Un diseño de la aplicación adaptativo, que se adapte a las diferentes pantallas de dispositivos.
Prioridad:	Alto

Nota: Elaboración propia

La Tabla 16 describe el Requerimiento No Funcional 02, relacionado con la usabilidad del sistema. Este requisito asegura que la interfaz sea amigable y fácil de utilizar, facilitando la interacción de los usuarios tanto en la carga de datos como en la interpretación de los resultados.

Tabla 16

Requerimiento No funcional 02 Usabilidad

Código:	RNF02
Nombre:	Usabilidad
Descripción:	La interfaz de usuario debe ser amigable y fácil de usar, facilitando la interacción tanto para la carga de datos como para la interpretación de resultados.
Prioridad:	Alto

Nota: Elaboración propia

La Tabla 17 muestra el Requerimiento No Funcional 03, enfocado en el rendimiento del sistema. Este requerimiento establece que la aplicación debe ser capaz de gestionar grandes volúmenes de datos de manera eficiente.

Tabla 17

Requerimiento No funcional 03 Rendimiento

Código:	RNF03
Nombre:	Rendimiento
Descripción:	La aplicación debe poder manejar cantidades grandes de datos de manera eficaz.
Prioridad:	Alto

Nota: Elaboración propia

La Tabla 19 detalla el Requerimiento No Funcional 04, que aborda la seguridad del sistema. Este requisito busca asegurar que se implementen mecanismos adecuados de autenticación y autorización para proteger el acceso a la aplicación y garantizar la seguridad de los datos.

Tabla 18

Requerimiento No funcional 04 Seguridad

Código:	RNF04
Nombre:	Seguridad
Descripción:	Garantizar la autenticación y autorización adecuadas para acceder a la aplicación.
Prioridad:	Alto

Fuente: Elaboración propia

3.2.1.7 Historias de Usuario

Las historias de usuario presentadas en la Tabla 19 sirven para ilustrar las interacciones entre el usuario y la aplicación.

Tabla 19

Historias de Usuario para la Limpieza y Preparación de Datos

Funcionalidad	Historia de Usuario
Recopilación de Datos	**Como** Administrador **quiero** subir los datos de las encuestas **para** procesarlos.
Procesamiento de Datos	**Como** Administrador **quiero** procesar los datos de las encuestas **para** crear un nuevo dataset de calidad.
Selección de Datos Relevantes	**Como** Administrador **quiero** seleccionar los datos relevantes **para** obtener un conjunto de variables identificadas.
Transformación de Datos	**Como** Administrador **quiero** transformar los datos de las encuestas de egresados **para** obtener datos adecuados para las operaciones de minería de datos.
Visualización de Resultados	**Como** Administrador **quiero** un informe de los datos recopilados **para** visualizar de manera clara y comprensible los resultados del proceso KDD.
Exportación de Resultados	**Como** Administrador **quiero** exportar los resultados **para** su posterior análisis o presentación.

Nota: Elaboración propia.

3.2.1.8 Tecnologías Utilizadas en el Desarrollo del Sistema

En esta sección se detallan las tecnologías empleadas para el desarrollo y operación del sistema de limpieza y preparación de datos. La Tabla 14 describen cada tecnología utilizada y su aplicación dentro de la plataforma.

Tabla 20

Especificación de Tecnologías Utilizadas en el Proyecto

Tecnología	Descripción	Uso
Python	Lenguaje de programación para el desarrollo de la aplicación.	Desarrollar la lógica del servidor, manejar peticiones HTTP, procesar datos, y generar archivos (PDF/Excel).
Django	Framework web de alto nivel en Python.	Construir la estructura de la aplicación web, gestionar rutas, vistas, modelos y base de datos.
Bootstrap	Framework de CSS, HTML y JavaScript para interfaces modernas y responsivas.	Diseñar la interfaz gráfica (botones, formularios, tablas, etc.).
HTML/CSS	HTML estructura las páginas web, CSS les da estilo y diseño visual.	Estructurar el contenido y aplicar estilos para un diseño atractivo y funcional.
JavaScript	Lenguaje de programación para añadir interactividad en páginas web.	Crear funciones dinámicas como validación de formularios y manejo de eventos.
Matplotlib	Biblioteca de Python para crear gráficos y visualizaciones.	Generar gráficos y visualizaciones en reportes (Excel y PDF).
Seaborn	Biblioteca basada en Matplotlib para gráficos estadísticos.	Crear gráficos como countplots y violinplots en los reportes.
Pandas	Biblioteca de Python para procesamiento y análisis de datos.	Procesar datos de bases de datos, facilitando la agrupación y análisis.
Openpyxl	Biblioteca para trabajar con archivos Excel en Python.	Generar, manipular y guardar archivos Excel con resultados de análisis de datos.
ReportLab	Biblioteca para crear documentos PDF en Python.	Generar reportes PDF con textos, tablas y gráficos.
Scikit-learn	Biblioteca de aprendizaje automático para Python.	Realizar agrupaciones y otras técnicas de análisis de datos.

Tecnología	Descripción	Uso
SQLite	Sistema de gestión de bases de datos ligero, por defecto en Django.	Almacenar y gestionar datos necesarios, como información de usuarios y respuestas.

Nota: Elaboración propia.

3.2.2 Diseño del Sistema

La fase de diseño está alineada con el segundo objetivo específico del proyecto, que es diseñar el módulo de limpieza de datos que incluya funcionalidades para importar los datos, identificar y corregir inconsistencias, duplicados y datos faltantes. En esta fase se define la estructura general de la aplicación, la arquitectura de los módulos y la disposición de la interfaz de usuario para garantizar un flujo intuitivo y eficiente en el manejo y procesamiento de la información.

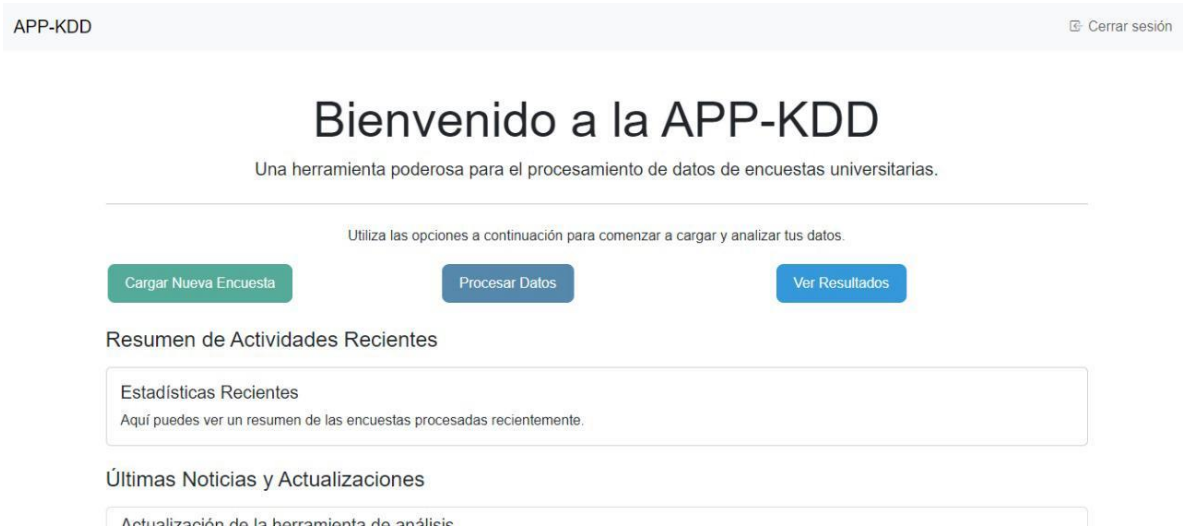
3.2.2.1 Diseño del Prototipo de la Interfaz

El prototipo de la interfaz de la plataforma web para la limpieza y preparación de datos se desarrolló en Figma, proporcionando una vista preliminar de la versión final de la interfaz gráfica y aprovechando sus herramientas y elementos para un diseño preciso.

En la Figura 7 se muestra el prototipo de la interfaz correspondiente al panel inicial de la aplicación. Este diseño organiza las funcionalidades principales y ofrece un acceso directo a las diferentes secciones de la plataforma, facilitando la navegación y el uso de la herramienta.

Figura 7

Prototipo de panel inicial de la aplicación

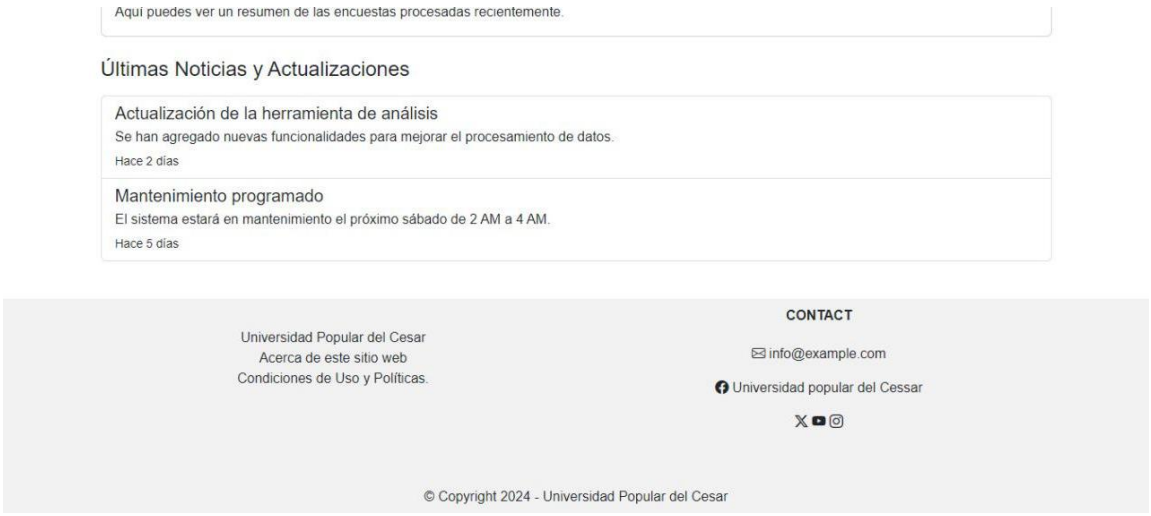


Nota. Elaboración propia

La Figura 8 presenta el prototipo de la interfaz para la sección de actualizaciones. Esta vista permite a los usuarios revisar y gestionar las últimas modificaciones realizadas en el sistema, manteniendo un registro actualizado de los cambios.

Figura 8

Prototipo de vista de actualizaciones del sistema.

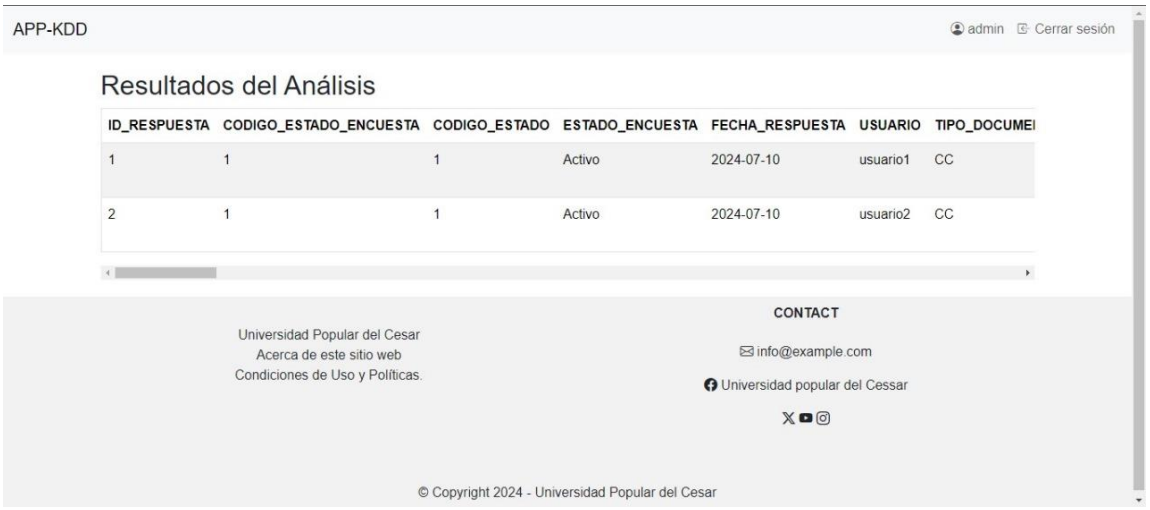


Nota. Elaboración propia

En la Figura 9 se observa la interfaz preliminar de los resultados del análisis, donde se muestran gráficos y tablas que permiten a los usuarios visualizar de manera clara el estado de los datos procesados, destacando patrones y tendencias relevantes.

Figura 9

Prototipo de resultados del análisis de datos

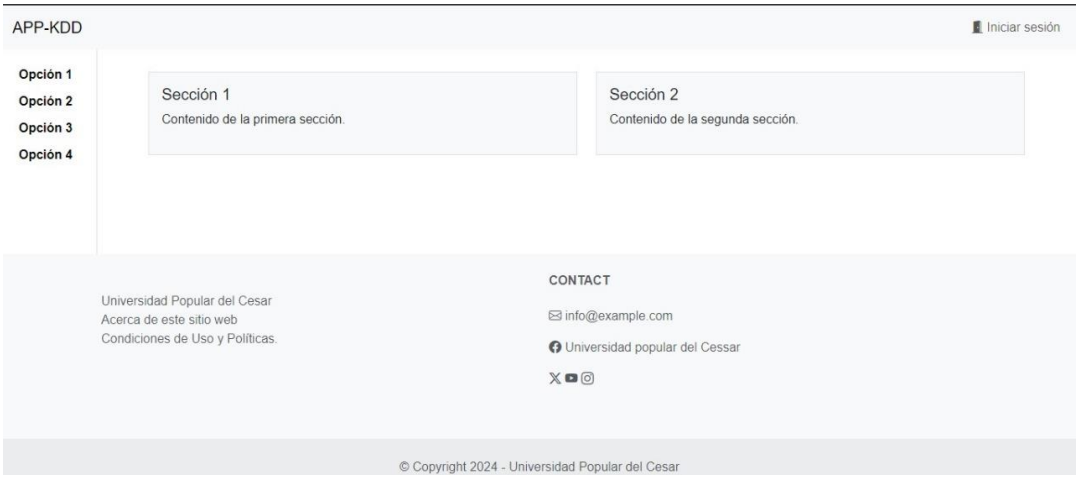


Nota. Elaboración propia

La Figura 10 ilustra la propuesta para visualizar las diferentes secciones de la aplicación. Aquí se destacan los módulos y funcionalidades organizados en pestañas o secciones que facilitan el acceso a cada una de las etapas del proceso de preparación y limpieza de datos.

Figura 10

Prototipo de navegación entre secciones de la aplicación

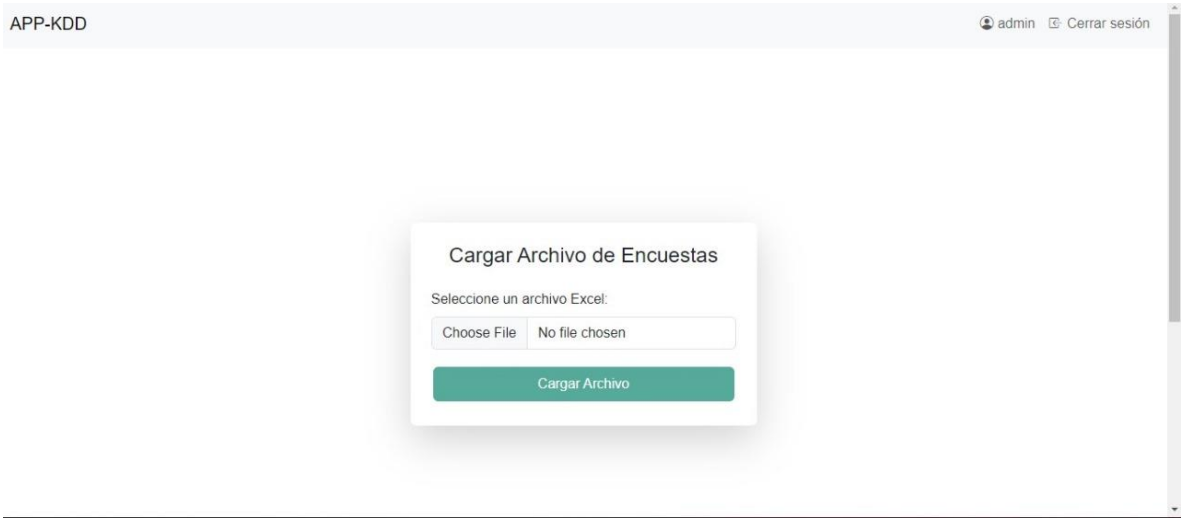


Nota. Elaboración propia

La Figura 11 muestra el modelo de la vista de carga de archivos, donde los usuarios pueden subir datos en formatos como CSV o Excel. Esta interfaz proporciona indicaciones claras para la carga de archivos y la validación de estos antes de ser procesados.

Figura 11

Prototipo de selección de encuesta para análisis

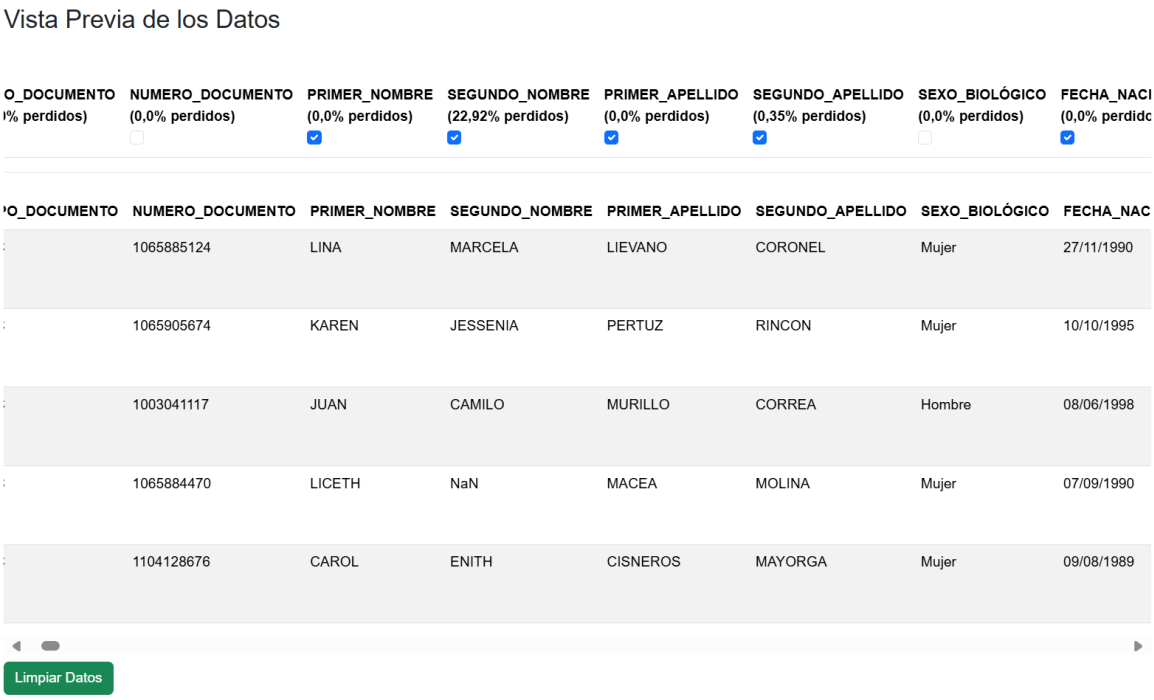


Nota. Elaboración propia

La Figura 12 muestra el prototipo del módulo de limpieza de datos. Aquí se pueden identificar y corregir inconsistencias, duplicados y datos faltantes, permitiendo asegurar la calidad de la información antes de su análisis.

Figura 12

Prototipo de módulo de limpieza y corrección de datos.

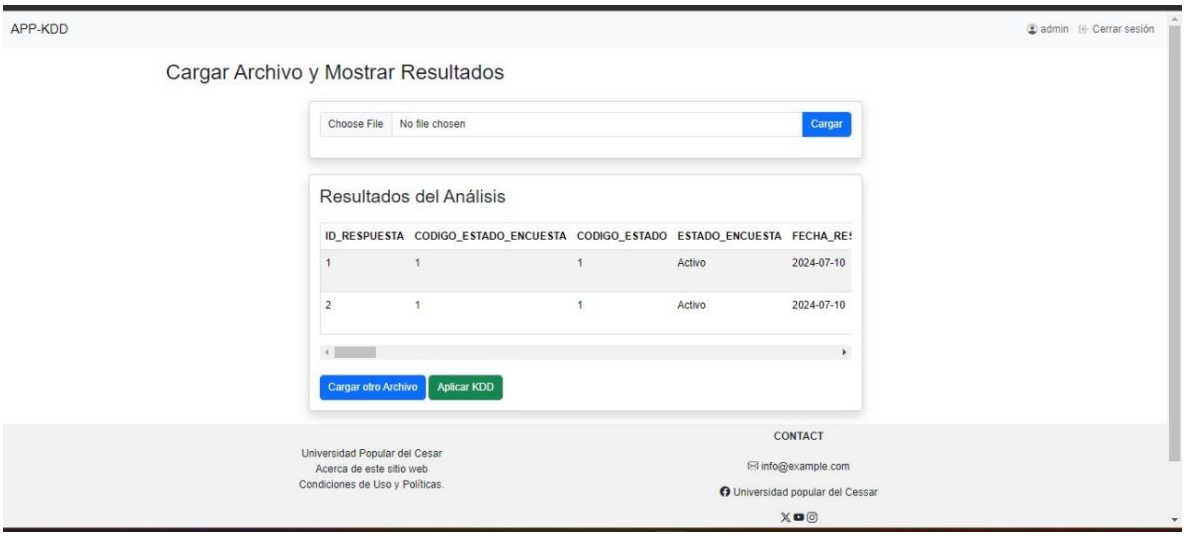


Nota. Elaboración propia

Finalmente, en la Figura 13 se presenta el prototipo de la vista de resultados, donde los usuarios pueden ver un resumen de los datos procesados y la efectividad de las operaciones realizadas, facilitando la generación de reportes finales.

Figura 13

Prototipo de resumen final de resultados procesados.



Nota. Elaboración propia

Este subcapítulo describe la creación de los módulos de la aplicación web, orientados a automatizar las etapas iniciales del proceso KDD (Knowledge Discovery in Databases) sobre los datos recolectados en las encuestas a egresados. La implementación de estos módulos se encuentra alineada con el tercer objetivo del proyecto, el cual busca crear los módulos de la aplicación web, incorporando las funcionalidades necesarias para la automatización del proceso KDD, optimizando así la preparación, limpieza y análisis de la información de manera eficiente y sistemática.

3.2.3 Desarrollo de los Módulos del sistema

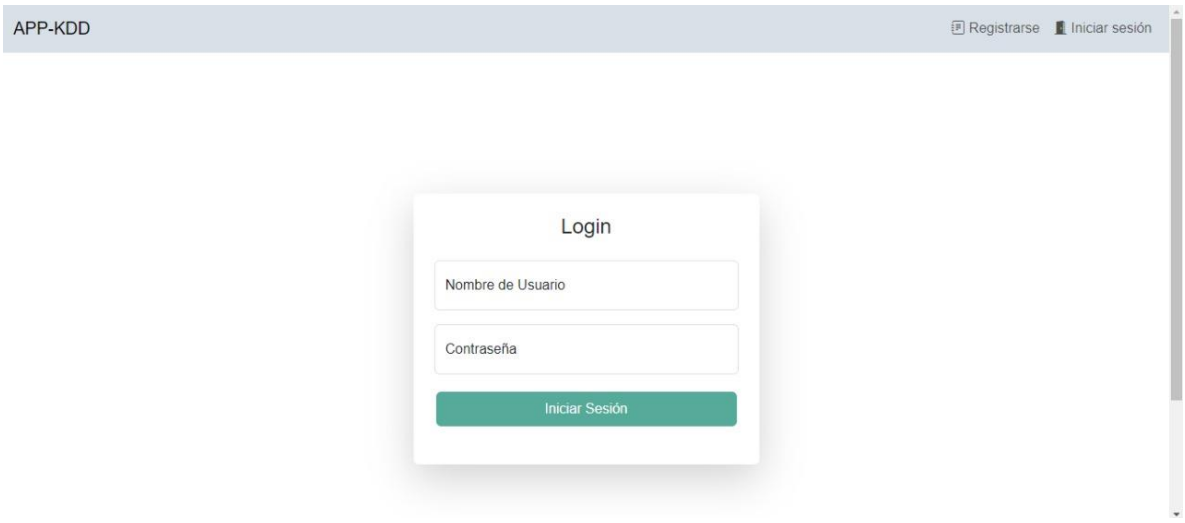
Este subcapítulo se enfoca en la creación de los módulos de la aplicación web, esto alineado con el tercer objetivo del proyecto, que consiste en desarrollar las funcionalidades para automatizar las primeras fases del proceso KDD y facilitar con ello la visualización y exportación de los datos con óptima calidad.

A continuación, se presentan los diferentes módulos y funcionalidades ya codificados y operativos.

En la Figura 14 se observa la vista de inicio de sesión de la aplicación, donde los usuarios deben ingresar sus credenciales para acceder al sistema. Esta es la versión final, que garantiza la seguridad del acceso a la plataforma.

Figura 14

Interfaz de inicio de sesión



Nota. Elaboración propia

En la Figura 15 se muestra el formulario para la creación de nuevos usuarios, diseñado para que los administradores gestionen el acceso al sistema. Permite añadir nuevos miembros con información básica, como nombre, correo y rol dentro de la plataforma.

Figura 15

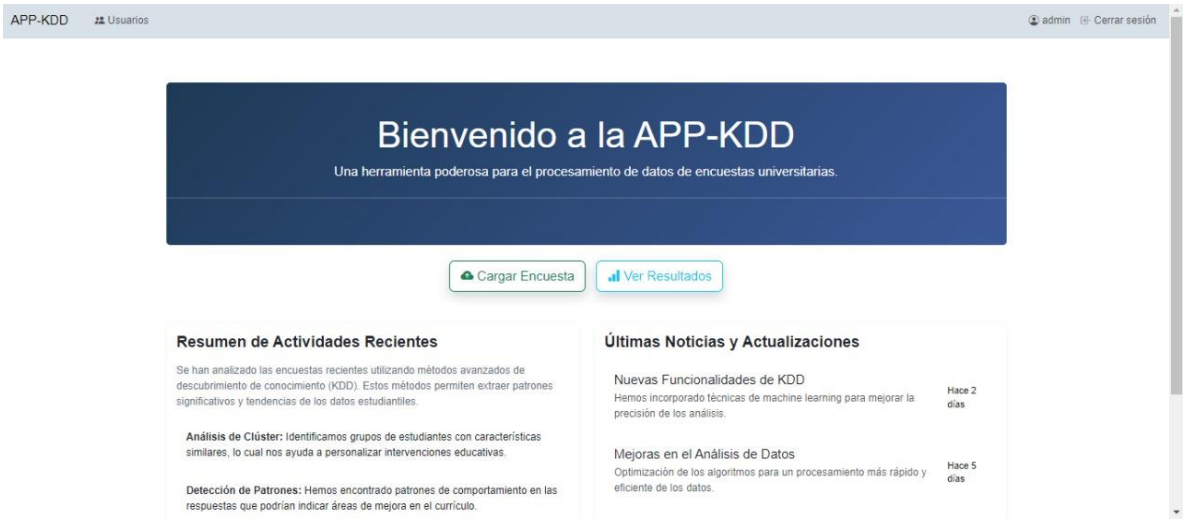
Registro de nuevos usuarios

Nota: Elaboración propia.

En la Figura 16 se aprecia la interfaz inicial del aplicativo, que actúa como panel de control central. Desde esta vista se puede acceder a las distintas opciones de la aplicación, como la carga de datos, limpieza y generación de reportes.

Figura 16

Panel principal de la aplicación

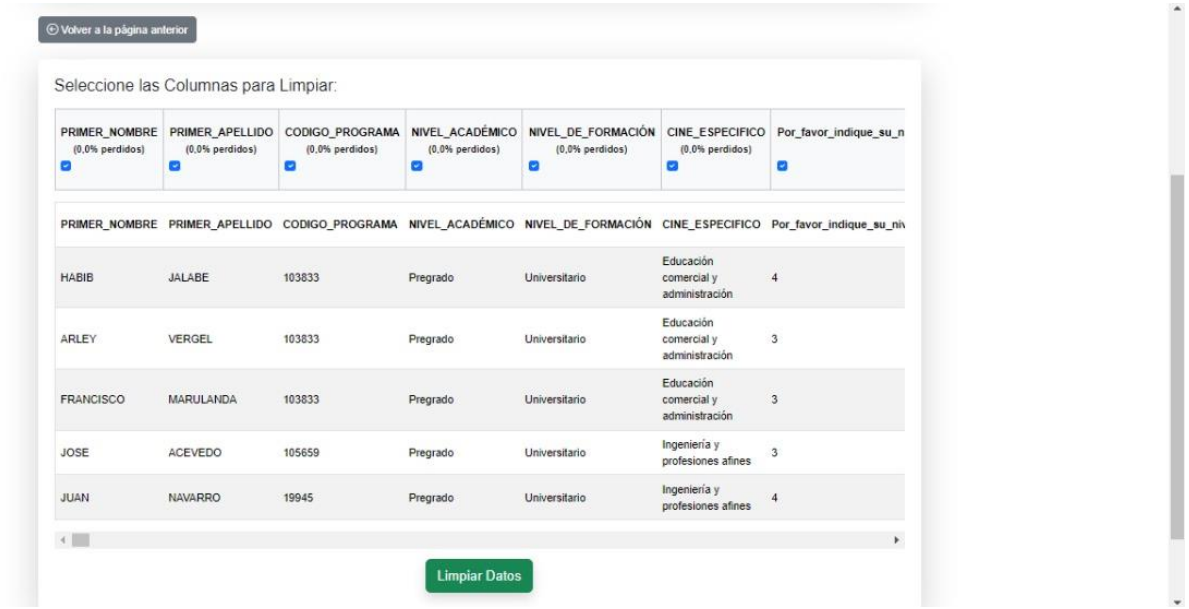


Nota: Elaboración propia.

En la imagen 17 se observa la vista de los datos importados en el sistema. Permite a los usuarios revisar y explorar los registros cargados antes de iniciar el proceso de limpieza, asegurando que la información esté correctamente estructurada.

Figura 17

Visualización de datos importados



Nota: Elaboración propia

En la imagen 18 se muestra el módulo de limpieza de datos, donde se identifican y corrigen las inconsistencias, duplicados y datos faltantes. Optimiza el proceso de depuración, asegurando la calidad de la información para su análisis.

Figura 18

Módulo de limpieza de datos

PRIMER_NOMBRE	PRIMER_APELLIDO	CODIGO_PROGRAMA	NIVEL_ACADÉMICO	NIVEL_DE_FORMACIÓN	CINE_ESPECIFICO	Por_favc
habib	jalabe	103833	pregrado	universitario	educación comercial y administración	4
arley	vergel	103833	pregrado	universitario	educación comercial y administración	3
francisco	marulanda	103833	pregrado	universitario	educación comercial y administración	3
jose	acevedo	105659	pregrado	universitario	ingeniería y profesiones afines	3

Nota: Elaboración propia

En la imagen 19 se presenta la vista final de los resultados del proceso de limpieza de datos. Aquí se visualizan gráficos y tablas generados a partir de los datos depurados, brindando un resumen visual y detallado para facilitar la interpretación de los resultados.

Figura 19

Visualización de resultados finales

Nombre	Apellido	Respuesta
yiseth	sanchez	1
yasbleydis	polo	1
angie	rodriguez	1
raquel	pallares	1
brandon	velasco	2
jahir	ojeda	2
michell	almanza	2
katty	parra	2
danilo	vergel	2
elias	sanchez	2

Nota: Elaboración propia

3.2.3 Implementación de la Aplicación en Entorno de Producción y Evaluación del Funcionamiento

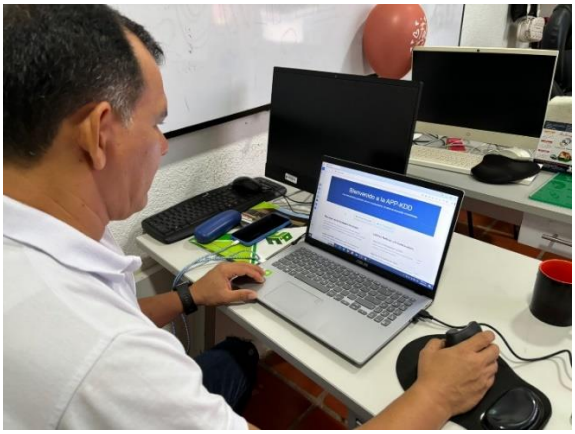
Este subcapítulo detalla la implementación de la aplicación web en un entorno de producción, en cumplimiento del cuarto objetivo del proyecto, que busca asegurar la eficiente preparación y limpieza de los datos de las encuestas de egresados. La implementación garantiza que la plataforma esté lista para su uso en un ambiente real, optimizando la gestión automatizada y el análisis de la información en un entorno seguro y controlado. Se incluyen los procesos de despliegue, configuración del servidor y pruebas funcionales realizadas para verificar el correcto funcionamiento de los módulos y la integridad de los datos procesados.

Como parte de esta fase, se llevó a cabo una prueba piloto en la Oficina de Egresados para evaluar la operatividad integral de la plataforma para el proceso de KDD, evidencia de esta actividad se aprecia en la Figura 21. Durante la prueba, se brindó una capacitación detallada al equipo, quienes participaron activamente en el uso de las funcionalidades de la aplicación. La evaluación se apoyó en un Checklist exhaustivo que permitió validar cada módulo y documentar áreas que requerían ajustes.

La retroalimentación obtenida, junto con los resultados del Checklist (ver Anexo A), proporcionó información valiosa para optimizar la experiencia del usuario y asegurar que la aplicación web cumpliera con todos los requisitos y expectativas, garantizando su confiabilidad y precisión las tareas iniciales del proceso KDD.

Figura 20

Despliegue y puesta en producción en la Oficina de Egresados de la Seccional Aguachica



Nota: Fuente propia.

3.4 CONCLUSIONES

La implementación del proyecto para automatizar fases iniciales del proceso de KDD (Knowledge Discovery in Databases) ha logrado modernizar significativamente la gestión de datos de los egresados en la Universidad Popular del Cesar, Seccional Aguachica. A través del uso de metodologías ágiles y herramientas de vanguardia, se desarrolló una aplicación web que satisface los requerimientos establecidos, proporcionando una plataforma robusta y eficiente para la recolección, almacenamiento, limpieza y preparación de datos.

El análisis preliminar permitió identificar las necesidades y desafíos específicos de la Oficina de Egresados, priorizando la automatización de la limpieza y preparación de datos para asegurar la integridad, seguridad y accesibilidad de la información en futuros análisis. Las pruebas realizadas en un entorno de producción confirmaron la funcionalidad óptima de la aplicación, obteniendo resultados favorables y aceptación por parte del personal administrativo y los usuarios clave.

Esta solución tecnológica no solo optimiza la gestión de datos, sino que también proporciona una herramienta intuitiva y accesible, facilitando el acceso a información precisa y actualizada. Además, su desarrollo refuerza el compromiso de la universidad con la mejora continua de los servicios a los egresados, fortaleciendo el vínculo institucional y estableciendo un precedente para futuras innovaciones en la gestión de la información.

3.5 RECOMENDACIONES

Capacitar al personal de la Oficina de Egresados en el uso y gestión de la nueva aplicación web, incluyendo el manejo de herramientas de análisis de datos para garantizar un aprovechamiento completo de todas las funcionalidades de la plataforma. Es fundamental desarrollar manuales o instructivos interactivos que faciliten el aprendizaje continuo, así como material audiovisual que complemente las sesiones de capacitación, asegurando una adaptación fluida y eficiente. Además, se sugiere implementar una plataforma de soporte y actualización de conocimientos, que permita a los usuarios resolver dudas y mantenerse al día con las mejoras del sistema.

3.6 REFERENCIAS BIBLIOGRÁFICAS

[1] W. Y. Ayele, "Adapting CRISP-DM for Idea Mining: A Data Mining Process for Generating Ideas Using a Textual Dataset," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, 2020, doi: https://doi.org/10.14569/IJACSA.2020.0110603.

[2] D. Maximini, *The Scrum Culture : Introducing Agile Methods in Organizations*. Cham, SWITZERLAND: Springer International Publishing AG, 2015. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=1966903

[3] J. López Ruiz and G. Garcia Brustenga, "Aplicación de un modelo de analítica de texto para la detección y clasificación automática de tendencias de investigación en e-Learning (caso UOC)," *Universitat Oberta de Catalunya (UOC)*, Nov. 2019, doi: 10.7238/elc.tendenciasinvestigacion.2019.

[4] R. Ordoñez, N. Salgado, J. Meza, and S. Ventura, "Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review," *Heliyon*, vol. 9, no. 3, Mar. 2023, doi: 10.1016/j.heliyon.2023.e13939.

[5] J. Cabrera, "Modelo para la identificación de relaciones entre la información sobre los graduados de los programas de Maestría y Doctorado de la Universidad Nacional de Colombia y su tiempo de permanencia," 2018. [Online]. Available: https://repositorio.unal.edu.co/handle/unal/62974

[6] M. Ropero, "Mejorar el modelo de estimación de riesgo de deserción de los estudiantes de pregrado de la Universidad Autónoma de Bucaramanga empleando herramientas business intelligence soportadas en software libre," 2018. [Online]. Available: http://hdl.handle.net/20.500.12749/3439

[7] A. Rodríguez Vasquez and M. A. Ochoa Ariza, "Relación entre el proceso de orientación vocacional implementado en las instituciones de educación media del departamento del Cesar y la deserción estudiantil en la Universidad Popular del Cesar," 2012. [Online]. Available: http://repositorio.unimagdalena.edu.co/handle/123456789/1517

[8] B. Chindoy and K. Diaz, "Desarrollo de un software que permita la detección de estudiantes desertores del programa de Ingeniería de Sistemas utilizando procesos y elementos de minería de datos," 2019. Accessed: Apr. 15, 2024. [Online]. Available: https://repositorioinstitucional.ufpso.edu.co/xmlui/handle/20.500.14167/1258

[9] F. Geerts, G. Mecca, P. Papotti, and D. Santoro, “Cleaning data with Llunatic.,” *VLDB Journal International Journal on Very Large Data Bases*, vol. 29, no. 4, pp. 867–892, Jul. 2020, doi: 10.1007/s00778-019-00586-5.

[10] O. Campesato, *Data Cleaning Pocket Primer*. Bloomfield, UNITED STATES: Mercury Learning & Information, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=30331791

[11] I. F. Ilyas and X. Chu, *Data Cleaning*. San Rafael, UNITED STATES: Association for Computing Machinery, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6954840

[12] J. K. Pilowsky, R. Elliott, and M. A. Roche, “Data cleaning for clinician researchers: Application and explanation of a data-quality framework,” *Australian Critical Care*, vol. 37, no. 5, pp. 827–833, 2024, doi: https://doi.org/10.1016/j.aucc.2024.03.004.

[13] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining : Practical Machine Learning Tools and Techniques*. San Diego, UNITED STATES: Elsevier Science & Technology, 2011. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=634862

[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Diego, UNITED STATES: Elsevier Science & Technology, 2011. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=729031

[15] L. Rokach and O. Z. Maimon, *Data Mining With Decision Trees: Theory And Applications*. Singapore, SINGAPORE: World Scientific Publishing Company, 2007. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=1679477

[16] J. L. M. Galiano, *Implantar SCRUM Con éxito*. Barcelona, SPAIN: Editorial UOC, 2016. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=7051300

[17] P. Li, *JIRA Agile Essentials : Bring the Power of Agile to Atlassian JIRA and Run Your Projects Efficiently with Scrum and Kanban*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2015. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=2077522

[18] N. Bäumer, *Mit strukturierter Agilität zu außergewöhnlichen Ideen : Wenn aus Scrum murcS wird*. Göttingen, GERMANY: BusinessVillage, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6226375

[19] J. Bell, *Machine Learning : Hands-On for Developers and Technical Professionals*. Somerset, UNITED STATES: John Wiley & Sons, Incorporated, 2014. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=1818248

[20] U. N. Dulhare, K. Ahmad, and K. A. Bin Ahmad, *Machine Learning and Big Data : Concepts, Algorithms, Tools and Applications*. Newark, UNITED STATES: John Wiley & Sons, Incorporated, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6268187

[21] R. Gopalakrishnan and A. Venkateswarlu, *Machine Learning for Mobile : Practical Guide to Building Intelligent Mobile Applications Powered by Machine Learning*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5628277

[22] M. A. Jabbar, *Machine Learning Methods for Signal, Image and Speech Processing*. Aalborg, DENMARK: River Publishers, 2021. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=29002971

[23] W.-M. Lee, *Python Machine Learning*. Newark, UNITED STATES: John Wiley & Sons, Incorporated, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5747364

[24] Z. Nagy, *Artificial Intelligence and Machine Learning Fundamentals : Develop Real-World Applications Powered by the Latest AI Advances*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2018. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5620491

[25] E. Jurczenko, *Machine Learning for Asset Management : New Developments and Financial Applications*. Newark, UNITED STATES: John Wiley & Sons, Incorporated, 2020. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6268186

[26] G. Kyriakides and K. G. Margaritis, *Hands-On Ensemble Learning with Python : Build Highly Optimized Ensemble Machine Learning Models Using Scikit-Learn and Keras*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited, 2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5837325

[27] K. Ramasubramanian and J. Moolayil, *Applied Supervised Learning with R : Use Machine Learning Libraries of R to Build Models That Solve Business Problems and Predict Future Trends*. Birmingham, UNITED KINGDOM: Packt Publishing, Limited,

2019. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5784240

[28] V. Rajlich, *Software Engineering : The Current Practice*. Milton, UNITED KINGDOM: CRC Press LLC, 2011. [Online]. Available: http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=1581463

[29] M. D. Kadenic, K. Koumaditis, and L. Junker-Jensen, “Mastering scrum with a focus on team maturity and key components of scrum,” *Inf Softw Technol*, vol. 153, p. 107079, 2023, doi: https://doi.org/10.1016/j.infsof.2022.107079.

[30] P. Becker, G. Tebes, D. Peppino, and L. Olsina, “Applying an Improving Strategy that embeds Functional and Non-Functional Requirements Concepts.,” *Journal of Computer Science & Technology (JCS&T)*, vol. 19, no. 2, pp. 153–174, Oct. 2019, doi: 10.24215/16666038.19.e15.

4. ANEXOS

ANEXO A. Checklist de funcionamiento en producción de la aplicación web

CHECKLIST PARA LA VERIFICACIÓN DE LA APLICACIÓN KDD

1. Acceso y Configuración Inicial

- [ ] **Acceso a la Aplicación**: Verifique que puede acceder a la aplicación utilizando sus credenciales proporcionadas.
- [ ] **Carga de Archivos XLSX**: Intente cargar un archivo XLSX y asegúrese de que la aplicación lo acepte sin errores.

2. Recopilación y Procesamiento de Datos

- **Recopilación de Datos (RF01)**:
  - [ ] Verifique que puede cargar correctamente archivos XLSX con los datos de encuestas.
  - [ ] Asegúrese de que el sistema muestra una confirmación de carga exitosa o un mensaje de error en caso de problemas.
- **Procesamiento de Calidad de Datos (RF02)**:
  - [ ] Asegúrese de que la aplicación identifica y aplica correcciones para valores faltantes o erróneos.
  - [ ] Compruebe que se pueden aplicar las correcciones propuestas (por ejemplo, permutación por moda o media).

3. Análisis y Selección de Datos

- **Selección de Datos Relevantes (RF03)**:
  - [ ] Verifique que puede ajustar los parámetros para influir en la selección de estas variables.
- **Transformación de Datos para Minería (RF04)**:
  - [ ] Verifique que los datos pueden ser transformados (normalización, estandarización) según lo requerido para el análisis.
  - [ ] Asegúrese de que los datos transformados están listos para ser utilizados en minería de datos.

4. Visualización y Exportación de Resultados

- **Visualización de Resultados (RF05):**
  - Asegúrese de que los gráficos y resúmenes visuales se muestran correctamente en todas las secciones de la aplicación.
  - Compruebe que los gráficos son claros y comprensibles, destacando patrones y tendencias importantes.
- **Exportación de Resultados (RF06):**
  - Verifique que puede exportar los resultados en formatos PDF.
  - Asegúrese de que la exportación respeta los filtros y configuraciones aplicados.

5. Usabilidad y Experiencia de Usuario

- **Accesibilidad (RNF01):**
  - Asegúrese de que la aplicación se visualiza correctamente.
  - Verifique que todos los elementos de la interfaz de usuario son accesibles y funcionan como se espera.
- **Usabilidad (RNF02):**
  - Asegúrese de que la interfaz de usuario es intuitiva y fácil de usar.
  - Verifique que los mensajes de error y confirmación son claros y útiles.

6. Rendimiento y Seguridad

- ***Rendimiento (RNF03):***
  - Verifique que la aplicación maneja eficientemente los datos cargados, sin tiempos de respuesta lentos (dependiendo de las especificaciones del equipo donde se implemente).
  - Asegúrese de que la aplicación no se bloquea o ralentiza durante el uso (dependiendo de las especificaciones del equipo donde se implemente).

- **Seguridad (RNF04):**
  - [x] Verifique que puede acceder solo a las funcionalidades para las que está autorizado.
  - [x] Asegúrese de que la aplicación requiere autenticación para acceder a áreas protegidas.

7. Evaluación Final

- **Cumplimiento de Objetivos:**
  - [x] Verifique que la aplicación cumple con todos los objetivos planeados para el procesamiento, análisis y visualización de los datos de egresados.
  - [x] Asegúrese de que todas las funcionalidades claves están disponibles y funcionan correctamente.
- **Feedback y Ajustes:**

Anote cualquier problema o dificultad que encuentre durante el uso.

Proporcione feedback al equipo de desarrollo para posibles mejoras o ajustes.
Generar reportes en archivo Excel ó word.

Realizado por:
José Gregorio Jorge García
C.C. 18.924.989

Nombre: