Prototipo de aplicativo de visualización de resultados de pruebas Saber 11 a nivel nacional de los años 2014 a 2023

John Fredy Leal

Asesor

Isaac Esteban Camargo Freile

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería - ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Nota de Aceptación

_____

Isaac Esteban Camargo Freile

Director de Trabajo de Grado

_____ _____

Jurado Jurado

2024

Resumen

La evaluación de los resultados del ICFES en Colombia es crucial para entender y mejorar la calidad educativa en el país. Sin embargo, los datos del ICFES, que abarcan múltiples años y contienen información detallada de estudiantes, colegios y municipios, están dispersos y no estructurados para un análisis fácil y comprensible. Los investigadores enfrentan dificultades para acceder, particionar y analizar estos datos de manera eficiente, lo que limita la capacidad para hacer comparaciones significativas y obtener insights valiosos.

Este proyecto pretende crear una herramienta que facilite a los investigadores el particionamiento, visualización y análisis de los datos de los resultados de las pruebas Saber 11 que practica el ICFES, con el fin de simplificar el proceso de análisis, generar comparaciones significativas y obtener información clave que apoye la toma de decisiones en el ámbito educativo a partir de los resultados obtenidos por los investigadores.

La funcionalidad del software brinda la posibilidad de que el usuario seleccione criterios de particionamiento de los datos sobre filas y columnas, generando nuevos conjuntos de datos más pequeños, ajustados a su necesidad, que posteriormente, podrán ser comparados por medio de las visualizaciones más comunes, se generarán estadísticas básicas de cada uno y comparativas, así como pruebas de normalidad e interpretación de los resultados por medio de un asistente de IA.

El proyecto se desarrollará durante 6 meses, con la metodología de desarrollo de software Scrum y el lenguaje de programación Python para el backend y TextScript para el frontend, con recursos propios y los resultados esperados son un prototipo con la mayoría de la funcionalidad propuesta (se excluyen temas de autenticación, suscripción, cobros, integración con otras plataformas) y una base de datos Mongo DB con las información limpia y organizada de los

resultados de las pruebas Saber 11 de los años 2014 a 2023, que incluya georeferenciación de los colegios e información de población de cada municipio. El proyecto está planteado para desarrollarse durante 6 meses con la metodología de desarrollo de software Scrum y el lenguaje de programación Python.

Palabras claves**:** Data analysis; data visualization; ICFES; python programming; open data.

Abstract

The evaluation of ICFES results in Colombia is crucial to understanding and improving the quality of education in the country. However, ICFES data, which spans multiple years and contains detailed information about students, schools, and municipalities, is scattered and unstructured, making it difficult to analyze in an easy and comprehensible way. Researchers face challenges in accessing, partitioning, and analyzing this data efficiently, limiting their ability to make meaningful comparisons and gain valuable insights.

This project aims to create a tool that facilitates the partitioning, visualization, and analysis of data from the Saber 11 test results conducted by ICFES, with the purpose of simplifying the analysis process, generating meaningful comparisons, and obtaining key information that supports decision-making in the educational field based on researchers' findings.

The software functionality will allow users to select data partitioning criteria for rows and columns, generating smaller, customized datasets that can subsequently be compared using common visualizations. Basic statistics for each dataset and comparative statistics will be generated, along with normality tests and AI-assisted interpretation of the results.

The project will be developed over six months, following the Scrum software development methodology, with Python as the backend programming language and TypeScript for the frontend.

The expected outcomes include a prototype with most of the proposed functionality (excluding authentication, subscriptions, payments, and integration with other platforms) and a MongoDB database containing clean and organized information from Saber 11 test results from 2014 to 2023, including georeferenced schools and population information for each municipality.

The project is planned for a six-month development timeline using the Scrum methodology and Python programming language.

***Keywords*:** Data Analysis; Data Visualization; ICFES; Python Programming; Open Data

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

El Examen de Estado de la Educación Media, conocido como Saber 11°, es una evaluación estandarizada que mide la calidad de la educación formal en Colombia para los estudiantes que terminan la educación media. Este examen está compuesto por cinco pruebas: Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés. (ICFES, 2024a). Esta prueba tiene como objetivos:

- Ser requisito obligatorio para el ingreso a la educación superior.
- Brindar información a los estudiantes sobre sus competencias en las diferentes áreas que sirva como apoyo para la orientación sobre su opción profesional.
- Ser criterio para la autoevaluación de los establecimientos educativos en función de sus proyectos educativos y planes de mejoramiento.
- Ser criterio para otorgar beneficios educativos (becas, premios).
- Ser la base para estudios de carácter cultural, social, económico y educativo y retroalimentar el quehacer de la evaluación.

La información sobre los resultados es publicada por el ICFES en diferentes plataformas, incluyendo el portal de datos abiertos del estado (MINTIC, 2024) y en la oficina de investigaciones del ICFES (ICFES, 2024b), correo electrónico para solicitudes de información del ICFES entre otros. Para poder acceder a estas fuentes es necesario tener usuarios registrados.

Los datos publicados por el ICFES sobre las pruebas saber 11, incluyen variables de caracterización de los colegios, variables de caracterización del estudiante y variables socioeconómicas. Teniendo en cuenta el último objetivo de la prueba y luego de realizar una revisión documental inicial sobre estudios realizados tomando como base los datos de resultados del ICFES tales como (Ferrer, 2010), (Bahamón & Reyes, 2014), por citar algunos, se determinó

que la mayoría, si no, todos ellos realizaron una limpieza de los datos y una selección de su grupo de interés, (Ferrer, 2010) solo se remite a los datos de Risaralda y a la variables de resultados en lenguaje, (Bahamón & Reyes, 2014), se limita solo al año 2012, (Rosero et al., 2021) se limitan al departamento de Nariño y al año 2018.

Una revisión inicial del portal de datos abiertos del ICFES arrojó que los sets de datos de diferentes periodos, o inclusive, del mismo periodo, pero publicados en diferente plataforma, tienen problemas tales como: Registros repetidos, diferentes conjuntos de variables, valores nulos entre otros. Esta situación lleva a concluir que una herramienta que permita realizar estas tareas repetitivas puede ayudar, disminuyendo tiempos y costos en la labor investigativa (López Murphy Y & Zarza, 2017). Aumentar la eficiencia en la elaboración de estudios puede ayudar a que más investigadores se interesen por estos sets de datos. Solamente en la UNAD, mientras realizaba este proyecto, al menos tres estudiantes tuvieron la necesidad de ubicar, descargar y limpiar datos de las pruebas saber 11. Se propone la elaboración de un aplicativo que cuente con los datos ya limpios y estandarizados de resultados de pruebas Saber 11 y que permita su segmentación y EDA (Exploratory Data Analysis) inicial, así como la visualización y cálculo de estadísticas básicas sobre las particiones de datos seleccionadas por el usuario.

Justificación

Desarrollar un sistema de información que centralice los datos de las pruebas saber 11 del ICFES y permita su análisis dinámico es esencial para mejorar la investigación educativa en Colombia. Este sistema facilitará a los investigadores el acceso a los datos, permitiéndoles realizar análisis detallados, comparar resultados entre diferentes años, colegios y municipios, y generar visualizaciones y mapas que ayuden en la toma de decisiones durante el proceso de preparación de los datos. Además, la automatización de estas tareas reducirá el tiempo y esfuerzo necesarios, aumentando la eficiencia de los estudios educativos.

La propuesta de desarrollar un aplicativo que integre la limpieza, estandarización y análisis exploratorio de los resultados de las pruebas Saber 11 busca abordar los desafíos identificados en la revisión inicial del portal de datos abiertos del ICFES. La creación de una herramienta que automatice estos procesos no solo aumentará la eficiencia en la preparación de datos, sino que también mejorará la calidad de los análisis y reducirá los costos asociados (López Murphy Y & Zarza, 2017). Esta solución facilitará el acceso a datos limpios y estandarizados, promoviendo el uso de estos conjuntos de datos en la investigación educativa y estimulando un mayor interés en el análisis de los resultados de las pruebas Saber 11, evitando el tedioso paso de la búsqueda y limpieza de estos.

Los estudios que puede apoyar este tipo de aplicativo incluyen, por ejemplo, la identificación de factores clave que afectan el rendimiento estudiantil tales como el acceso a internet, nivel socioeconómico o ubicación geográfica; evaluación del impacto de programas, realizando análisis comparativos antes y después de la aplicación del programa; diseño de programas personalizados para los estudiantes; evaluación del uso de tecnologías en el aprendizaje; identificación de desigualdades en el acceso al sistema educativo, entre otros.

Objetivos

Objetivo General

Desarrollar una aplicación web, utilizando herramientas de backend y frontend como Flask, React con TypeScript y TextScript para analizar y visualizar datos de las pruebas saber 11 del año 2014 a 2023, para la contribución de avance en la investigación educativa en Colombia.

Objetivos Específicos

Centralizar y organizar datos del ICFES: Consolidar los resultados de las pruebas Saber 11 en un único archivo CSV con las variables socioeconómicas, de resultados, georreferenciación de los colegios y datos de población de los municipios, accesible y optimizada para consultas eficientes mediante una API desarrollada en Flask, con opciones de exportación en formatos como CSV, JSON y Excel.

Facilitar la segmentación y filtrado de datos: Implementar funcionalidades en un frontend intuitivo construido con React y TextScript, que permitan a los usuarios segmentar conjuntos de datos de manera flexible utilizando variables específicas del dataset.

Proveer herramientas de análisis y visualización: Desarrollar opciones para realizar visualizaciones univariadas, validar supuestos estadísticos (normalidad y homocedasticidad) y generar rankings basados en variables de puntaje, acompañados de gráficos y análisis interpretativos.

Habilitar la comparación de sets de datos: Incorporar funcionalidades que permitan comparar set de datos mediante pruebas paramétricas y no paramétricas, con resultados presentados en Boxplots y análisis con asistencia de inteligencia artificial.

Marco de Referencia

Estado del Arte

El análisis estadístico asistido por herramientas gráficas e inteligencia artificial (IA) ha experimentado un crecimiento significativo en diversos países, incluyendo Estados Unidos, Reino Unido, Alemania, y países de América Latina, como Colombia, Brasil y México. En estos contextos, las aplicaciones de análisis de datos se han vuelto esenciales para procesar grandes volúmenes de información en áreas como la educación, la salud y la investigación científica.

Plataformas como Tableau (*¿Qué Es Tableau?*, n.d.), Power BI (*Power BI - Data Visualization | Microsoft Power Platform*, n.d.), y Google Data Studio (*Data Studio Now Available as a Google Cloud Service | Google Cloud Blog*, n.d.) son ampliamente utilizadas en empresas e instituciones educativas para realizar análisis exploratorios de datos (EDA) y crear visualizaciones interactivas sin necesidad de conocimientos avanzados en programación. Además, herramientas como RapidMiner (*Getting Started with RapidMiner Studio - Altair RapidMiner Documentation*, n.d.) y KNIME (*Open for Innovation | KNIME*, n.d.) combinan técnicas de machine learning con análisis visual, lo que permite realizar análisis predictivos y prescriptivos basados en datos sin necesidad de codificación avanzada.

Estas herramientas ofrecen una gran cantidad de funcionalidades, pero uno de los mayores desafíos que enfrentan los investigadores que no cuentan con bases en programación es la dificultad para acceder a datasets estructurados y comprender los algoritmos de IA que pueden aplicar. Si bien plataformas como las mencionadas permiten generar visualizaciones a partir de bases de datos, las barreras técnicas para la personalización y el análisis profundo de los datos siguen siendo altas para los usuarios sin conocimientos técnicos. Este problema es particularmente relevante en el ámbito educativo, donde investigadores y docentes necesitan

realizar análisis detallados de resultados como los exámenes nacionales o internacionales, tales como el ICFES en Colombia o el SAT en Estados Unidos.

En Colombia, por ejemplo, los investigadores que analizan los resultados de pruebas estandarizadas como las pruebas Saber 11 enfrentan dificultades para acceder a los datos de manera estructurada y poder analizarlos de forma eficiente. Herramientas como las mencionadas anteriormente pueden ser útiles, pero el nivel de detalle que los investigadores buscan, como la capacidad para particionar datos según variables específicas (como municipio, colegio o año), sigue siendo una tarea que requiere un esfuerzo considerable cuando no se tienen conocimientos en programación o manejo avanzado de datos.

Además, si bien aplicaciones como SPSS (*Software IBM SPSS*, n.d.) y STATA (*Statistical Software for Data Science | Stata*, n.d.) son muy utilizadas para análisis estadísticos en muchos países, presentan limitaciones cuando se trata de integrar análisis visuales avanzados con técnicas de IA. Esto dificulta la generación de insights accionables y comprensibles para quienes no tienen experiencia en estas áreas, lo que a menudo lleva a interpretaciones incompletas o superficiales de los resultados.

Por otra parte, investigaciones recientes sugieren que el uso de herramientas como “R”, · Jupyter Notebooks (*Project Jupyter | Home*, n.d.) y bibliotecas como Matplotlib (*Matplotlib — Visualization with Python*, n.d.), Plotly (*Plotly Python Graphing Library*, n.d.), y Seaborn (Waskom, 2021) en combinación con algoritmos de machine learning ha facilitado la capacidad de los investigadores con experiencia en programación para realizar análisis complejos y automatizar la interpretación de resultados. Sin embargo, esto aún está fuera del alcance de muchos profesionales que no cuentan con conocimientos técnicos profundos. Para aquellos

investigadores que no tienen una base sólida en programación, acceder a estas capacidades sigue siendo un reto considerable.

Por esta razón, el desarrollo de aplicaciones que combinen la simplicidad de las interfaces gráficas con el poder de la IA para el análisis estadístico está en auge. Herramientas como AutoML (*AutoML | AutoML*, n.d.), y H2O.ai (H2O.ai, 2024), están diseñadas para automatizar el proceso de análisis y generar modelos predictivos de manera intuitiva, reduciendo así la brecha técnica entre investigadores y la tecnología. No obstante, estas herramientas también tienen sus limitaciones, ya que no siempre permiten una personalización completa del análisis, lo que podría ser necesario en investigaciones académicas específicas.

Aunque las herramientas actuales han avanzado considerablemente en la democratización del análisis de datos, los investigadores que no cuentan con formación en programación todavía enfrentan importantes barreras (Shang et al., 2019). En países como Colombia, el análisis de datos educativos sigue siendo un desafío que requiere aplicaciones especializadas que no solo centralicen los datos, sino que también faciliten el acceso, particionamiento y análisis profundo con visualizaciones intuitivas y asistencia basada en IA. El desarrollo de plataformas accesibles que integren estas funcionalidades es clave para permitir que un mayor número de investigadores pueda aprovechar todo el potencial de los datos en sus respectivas áreas de estudio.

Marco Legal

En Colombia, el marco normativo sobre datos abiertos está diseñado para promover la transparencia, el acceso a la información pública y la rendición de cuentas. A continuación, se presentan las principales leyes y políticas relacionadas:

Ley 1712 de 2014

La Ley 1712 de 2014, también conocida como la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional, tiene como objetivo regular el derecho de acceso a la información pública, los procedimientos para su ejercicio y garantía, y las excepciones a la publicidad de la información.

Principios Clave

- Máxima Publicidad: Toda información en posesión de un sujeto obligado es pública, salvo excepciones constitucionales o legales.
- Transparencia: La información en poder de los sujetos obligados se presume pública y debe ser accesible en los términos más amplios posibles.
- Buena Fe: Los sujetos obligados deben cumplir con las obligaciones derivadas del derecho de acceso a la información pública de manera honesta y leal.
- Facilitación: Los sujetos obligados deben facilitar el ejercicio del derecho de acceso a la información pública.
- No Discriminación: La información debe ser entregada a todas las personas en igualdad de condiciones.
- Gratuidad: El acceso a la información pública es gratuito, salvo el costo de reproducción de la información.
- Celeridad: Se busca la agilidad en el trámite y la gestión administrativa1.

- Eficacia: Se impone el logro de resultados mínimos en relación con las responsabilidades confiadas a los organismos estatales.
- Calidad de la Información: La información debe ser oportuna, objetiva, veraz, completa, reutilizable, procesable y estar disponible en formatos accesibles1.
- Divulgación Proactiva: Los sujetos obligados deben promover y generar una cultura de transparencia.

Ámbito de Aplicación

La ley se aplica a todas las entidades públicas, órganos, organismos y entidades estatales independientes o autónomos, personas naturales y jurídicas que presten función pública, partidos políticos, y entidades que administren recursos públicos.

Excepciones

La ley establece excepciones al acceso a la información pública cuando su divulgación pueda causar daño a derechos de personas naturales o jurídicas, o a intereses públicos como la defensa y seguridad nacional, la seguridad pública, las relaciones internacionales, entre otros (Congreso de Colombia, 2014).

Decreto 103 de 2015

El Decreto 103 de 2015 reglamenta parcialmente la Ley 1712 de 2014, conocida como la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional. Este decreto establece las directrices y procedimientos para garantizar el acceso a la información pública en Colombia, promoviendo la transparencia y la rendición de cuentas por parte de las entidades públicas. El decreto define el objeto y el ámbito de aplicación, especificando que las disposiciones son aplicables a los sujetos obligados mencionados en el artículo 5 de la Ley 1712

de 2014. Estos sujetos incluyen entidades públicas y privadas que manejan recursos públicos o prestan servicios públicos.

Se establecen estándares para la publicación de información en sitios web oficiales, incluyendo la información mínima requerida, el registro de activos de información, el índice de información clasificada y reservada, y el esquema de publicación de información. También se detallan los procedimientos para la publicación de información contractual y de adquisiciones.

El decreto incluye directrices para garantizar la accesibilidad de la información a personas con discapacidad y a grupos étnicos y culturales, asegurando que la información pública esté disponible en formatos alternativos y accesibles. Se especifican los medios idóneos para recibir solicitudes de información pública, los procedimientos para responder a estas solicitudes, y los costos asociados a la reproducción de la información. También se establece el principio de gratuidad, indicando que no se deben cobrar costos adicionales a los de reproducción de la información.

El decreto define las excepciones al derecho de acceso a la información pública, incluyendo información clasificada por razones de defensa y seguridad nacional, estabilidad macroeconómica y financiera, y datos personales sensibles. Se establecen directrices para la calificación y divulgación de información clasificada o reservada.

Se detallan los instrumentos necesarios para la gestión de la información pública, como el registro de activos de información, el índice de información clasificada y reservada, el esquema de publicación de información, y el programa de gestión documental.

El Decreto 103 de 2015 se relaciona con la publicación de resultados de pruebas ICFES en cuanto a la transparencia y el acceso a la información pública. Las entidades encargadas de la evaluación educativa, como el ICFES, están obligadas a publicar y divulgar información

relevante, incluyendo los resultados de las pruebas, de manera accesible y transparente. Esto asegura que los ciudadanos puedan acceder a esta información de manera oportuna, promoviendo la rendición de cuentas y la transparencia en el sector educativo (Presidencia de la República, 2015).

La Guía de Datos Abiertos de Colombia

La Guía de Datos Abiertos de Colombia (Ministerio de Tecnologías de la Información y las Comunicaciones, 2021) está dirigida a las entidades sujetas a la aplicación de la Ley 1712 de 2014, conocida como la Ley de Transparencia y del Derecho de Acceso a la Información Pública. Su propósito es proporcionar orientaciones y buenas prácticas para el desarrollo de estrategias de apertura y reutilización de datos abiertos en el país. Tiene como objetivo principal fomentar la transparencia, la participación ciudadana y la innovación a través del acceso y uso de datos abiertos. Busca que las entidades públicas publiquen datos de manera proactiva y que estos sean reutilizables por cualquier persona interesada. Se detallan las estrategias para la apertura de datos, incluyendo la identificación de conjuntos de datos prioritarios, la implementación de políticas de datos abiertos y la creación de planes de acción específicos. También se enfatiza la importancia de la colaboración entre diferentes entidades y sectores para maximizar el impacto de los datos abiertos. Adicionalmente, establece los estándares y formatos recomendados para la publicación de datos abiertos, asegurando que sean accesibles, comprensibles y reutilizables. Se incluyen directrices sobre cómo documentar los datos, cómo garantizar su calidad y cómo promover su uso a través de licencias abiertas. Se destacan los beneficios de los datos abiertos, tanto para el gobierno como para la sociedad en general. Estos incluyen la mejora de la transparencia y la rendición de cuentas, el fomento de la innovación y el desarrollo económico, y la promoción de la participación ciudadana y la colaboración.

La guía proporciona un marco para la implementación y el seguimiento de las iniciativas de datos abiertos, incluyendo la evaluación de su impacto y la identificación de áreas de mejora. También se ofrecen recomendaciones para la capacitación y el desarrollo de capacidades en las entidades públicas.

Ley 1581 de 2012

Aunque esta ley se centra en la protección de datos personales, es fundamental mencionarla en el contexto de datos abiertos.

La Ley 1581 de 2012 (Congreso de Colombia, 2012), también conocida como la Ley de Protección de Datos Personales, establece disposiciones generales para la protección de datos personales en Colombia. Su objetivo es desarrollar el derecho constitucional de todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos.

Tiene como objetivo garantizar que el tratamiento de datos personales se realice respetando los derechos, libertades y garantías constitucionales. Establece principios y disposiciones aplicables a los datos personales registrados en cualquier base de datos, tanto de entidades públicas como privadas.

La ley se aplica al tratamiento de datos personales efectuado en territorio colombiano o cuando el responsable o encargado del tratamiento no establecido en territorio nacional le sea aplicable la legislación colombiana en virtud de normas y tratados internacionales. Existen excepciones, como bases de datos mantenidas en un ámbito personal o doméstico, y aquellas relacionadas con la seguridad y defensa nacional.

La ley establece principios para el tratamiento de datos personales, incluyendo:

- Legalidad: El tratamiento debe sujetarse a lo establecido en la ley.

- Finalidad: El tratamiento debe obedecer a una finalidad legítima.
- Libertad: El tratamiento solo puede ejercerse con el consentimiento previo, expreso e informado del titular.
- Veracidad o Calidad: La información debe ser veraz, completa, exacta, actualizada, comprobable y comprensible.
- Transparencia: Se debe garantizar el derecho del titular a obtener información sobre sus datos.
- Acceso y Circulación Restringida: El tratamiento se sujeta a los límites derivados de la naturaleza de los datos personales.
- Seguridad: La información debe manejarse con medidas necesarias para otorgar seguridad.
- Confidencialidad: Se debe garantizar la reserva de la información.

Los titulares de los datos personales tienen derechos como conocer, actualizar y rectificar sus datos, solicitar prueba de la autorización otorgada, ser informados sobre el uso de sus datos, presentar quejas ante la Superintendencia de Industria y Comercio, revocar la autorización y acceder gratuitamente a sus datos.

Los responsables y encargados del tratamiento deben cumplir con deberes como garantizar el ejercicio del derecho de hábeas data, conservar la información bajo condiciones de seguridad, actualizar y rectificar la información, tramitar consultas y reclamos, y adoptar manuales internos de políticas y procedimientos.

La Superintendencia de Industria y Comercio es la autoridad encargada de la vigilancia y sanción en materia de protección de datos personales. Puede imponer sanciones como multas, suspensión de actividades y cierre de operaciones relacionadas con el tratamiento de datos.

La Ley 1581 de 2012 se relaciona con la publicación de resultados de pruebas ICFES en cuanto a la protección de datos personales. Las entidades encargadas de la evaluación educativa, como el ICFES, deben garantizar que el tratamiento de los datos personales de los estudiantes se realice conforme a los principios y disposiciones de la ley, asegurando la privacidad y seguridad de la información, por lo tanto, los datos publicados por el ICFES e incluidos en el aplicativo se encuentran debidamente anonimizados y pueden ser usados dentro de los marcos legales.

Desarrollo del Proyecto

Levantamiento y Limpieza de Datos

Durante esta etapa del proyecto se obtuvieron archivos del portal de datos abiertos del estado, del portal del ICFES y del DANE.

Los archivos se encontraron con diferentes formatos, separadores, orden de las columnas, nombres de las columnas, registros duplicados y resultados de las pruebas en blanco. Se crearon varios programas en lenguaje Python para seleccionar las variables que se encontraban en toda la serie de los conjuntos de datos más completos. Se adicionaron variables de georeferenciación de los colegios y de población de los municipios de ubicación de los colegios. Finalmente se llegó a un conjunto de datos con 5.541.529 observaciones (filas) y 101 variables organizadas como se describe a continuación.

Información del Estudiante

Tabla 1

Variables de Información del Estudiante

Nombre de la Variable	Descripción	Tipo
ESTU_COD_RESIDE_DEPTO	Código del departamento de residencia	Categórica
ESTU_COD_RESIDE_MCPIO	Código del municipio de residencia	Categórica
ESTU_CONSECUTIVO	Código consecutivo del estudiante	Numérica
ESTU_DEPTO_RESIDE	Departamento de residencia	Categórica
ESTU_ESTADOINVESTIGACION	Estado en la investigación	Categórica
ESTU_ESTUDIANTE	Si es estudiante o no	Categórica
ESTU_FECHANACIMIENTO	Fecha de nacimiento	Fecha
ESTU_GENERACION	Generación	Categórica
ESTU_GENERO	Género	Categórica
ESTU_INSE_INDIVIDUAL	INSE individual	Numérica
ESTU_LENGUANATIVA	Lengua nativa	Categórica

ESTU_MCPIO_RESIDE	Municipio de residencia	Categórica
ESTU_NACIONALIDAD	Nacionalidad	Categórica
ESTU_NSE_ESTABLECIMIENTO	NSE del establecimiento	Categórica
ESTU_NSE_INDIVIDUAL	NSE individual	Categórica
ESTU_PAIS_RESIDE	País de residencia	Categórica
ESTU_PRESENTACIONSABADO	Presentación el sábado	Categórica
ESTU_TIPODOCUMENTO	Tipo de documento	Categórica

Nota. Listado de las variables del conjunto de datos relativas a las características del estudiante

Información Familiar

Tabla 2

Variables de Información Familiar

Nombre de la Variable	Descripción	Tipo
FAMI_COMECARNEPESCADOHUEVO	Consumo de carne, pescado y huevo	Categórica
FAMI_COMECEREALFRUTOSLEGUMBRE	Consumo de cereales, frutas y legumbres	Categórica
FAMI_COMELECHEDERIVADOS	Consumo de leche y derivados	Categórica
FAMI_COMECARNEPESCADOHUEVO	Consumo de carne, pescado y huevo	Categórica
FAMI_CUARTOSHOGAR	Número de cuartos en el hogar	Numérica
FAMI_EDUCACIONMADRE	Nivel educativo de la madre	Categórica
FAMI_EDUCACIONPADRE	Nivel educativo del padre	Categórica
FAMI_ESTRATOVIVIENDA	Estrato de vivienda	Categórica
FAMI_NUMLIBROS	Número de libros en el hogar	Numérica
FAMI_PERSONASHOGAR	Número de personas en el hogar	Numérica
FAMI_SITUACIONECONOMICA	Situación económica	Categórica
FAMI_TIENEAUTOMOVIL	Si la familia tiene automóvil	Categórica
FAMI_TIENECOMPUTADOR	Si la familia tiene computador	Categórica

FAMI_TIENECONSOLAVIDEOJUEGOS	Si la familia tiene consola de videojuegos	Categórica
FAMI_TIENEHORNOMICROOGAS	Si la familia tiene horno microondas o de gas	Categórica
FAMI_TIENEINTERNET	Si la familia tiene internet	Categórica
FAMI_TIENELAVADORA	Si la familia tiene lavadora	Categórica
FAMI_TIENEMOTOCICLETA	Si la familia tiene motocicleta	Categórica
FAMI_TIENESERVICIOTV	Si la familia tiene servicio de TV	Categórica
FAMI_TRABAJOLABORMADRE	Trabajo laboral de la madre	Categórica
FAMI_TRABAJOLABORPADRE	Trabajo laboral del padre	Categórica

Nota. Listado de las variables del conjunto de datos relativas a las características de la familia

Hábitos del Estudiante

Tabla 3

Variables de Hábitos del Estudiante

Nombre de la Variable	Descripción	Tipo
ESTU_DEDICACIONLECTURADIARIA	Dedicación a la lectura diaria	Numérica
ESTU_DEDICACIONINTERNET	Dedicación al uso de internet	Numérica
ESTU_HORASSEMANATRABAJA	Horas de trabajo semanales	Numérica
ESTU_TIPOREMUNERACION	Tipo de remuneración	Categórica

Nota. Listado de las variables del conjunto de datos relativas a los hábitos del estudiante

Información del Colegio

Tabla 4

Variables Sobre Información del Colegio

Nombre de la Variable	Descripción	Tipo
COLE_AREA_UBICACION	Área de ubicación del colegio	Categórica
COLE_BILINGUE	Si el colegio es bilingüe	Categórica
COLE_CALENDARIO	Calendario del colegio	Categórica
COLE_CARACTER	Carácter del colegio	Categórica
COLE_COD_DANE_ESTABLECIMIENTO	Código DANE del establecimiento	Categórica
COLE_COD_DANE_SEDE	Código DANE de la sede	Categórica
COLE_COD_DEPTO_UBICACION	Código del departamento de ubicación	Categórica
COLE_COD_MCPIO_UBICACION	Código del municipio de ubicación	Categórica
COLE_CODIGO_ICFES	Código ICFES del colegio	Categórica
COLE_DEPTO_UBICACION	Departamento de ubicación del colegio	Categórica
COLE_GENERO	Género del colegio	Categórica
COLE_JORNADA	Jornada del colegio	Categórica
COLE_MCPIO_UBICACION	Municipio de ubicación del colegio	Categórica
COLE_NATURALEZA	Naturaleza del colegio	Categórica
COLE_NOMBRE_ESTABLECIMIENTO	Nombre del establecimiento	Categórica
COLE_NOMBRE_SEDE	Nombre de la sede	Categórica
COLE_SEDE_PRINCIPAL	Si es la sede principal	Categórica

Nota. Listado de las variables del conjunto de datos relativas al colegio

Presentación del Examen

Tabla 5

Variable Sobre la Presentación del Examen

Nombre de la Variable	Descripción	Tipo
ESTU_PRIVADO_LIBERTAD	Si está privado de libertad	Categórica
ESTU_COD_DEPTO_PRESENTACION	Código del departamento de presentación	Categórica
ESTU_COD_MCPIO_PRESENTACION	Código del municipio de presentación	Categórica
ESTU_DEPTO_PRESENTACION	Departamento de presentación	Categórica
ESTU_MCPIO_PRESENTACION	Municipio de presentación	Categórica

Nota. Listado de las variables del conjunto de datos relativas al colegio

Resultados de la Prueba

Tabla 6

Variable de Resultados de la Prueba

Nombre de la Variable	Descripción	Tipo
DESEMP_C_NATURALES	Desempeño en Ciencias Naturales	Numérica
DESEMP_INGLES	Desempeño en Inglés	Numérica
DESEMP_LECTURA_CRITICA	Desempeño en Lectura Crítica	Numérica
DESEMP_MATEMATICAS	Desempeño en Matemáticas	Numérica
DESEMP_SOCIALES_CIUDADANAS	Desempeño en Sociales y Ciudadanas	Numérica
PERCENTIL_C_NATURALES	Percentil en Ciencias Naturales	Numérica
PERCENTIL_ESPECIAL_GLOBAL	Percentil especial global	Numérica
PERCENTIL_GLOBAL	Percentil global	Numérica
PERCENTIL_INGLES	Percentil en Inglés	Numérica
PERCENTIL_LECTURA_CRITICA	Percentil en Lectura Crítica	Numérica

PERCENTIL_MATEMATICAS	Percentil en Matemáticas	Numérica
PERCENTIL_SOCIALES_CIUDADANAS	Percentil en Sociales y Ciudadanas	Numérica
PUNT_C_NATURALES	Puntaje en Ciencias Naturales	Numérica
PUNT_GLOBAL	Puntaje global	Numérica
PUNT_INGLES	Puntaje en Inglés	Numérica
PUNT_LECTURA_CRITICA	Puntaje en Lectura Crítica	Numérica
PUNT_MATEMATICAS	Puntaje en Matemáticas	Numérica
PUNT_SOCIALES_CIUDADANAS	Puntaje en Sociales y Ciudadanas	Numérica

Nota. Listado de las variables del conjunto de datos relativas a los resultados de la prueba

Georeferenciación y Otros

Tabla 7

Variables de Georeferenciación y Otros

Nombre de la Variable	Descripción	Tipo
SEMESTRE_EXAMEN	Semestre del examen	Categórica
NOMBRE_INS	Nombre de la institución	Categórica
CORREO_INS	Correo electrónico de la institución	Categórica
DIRECCION	Dirección de la institución	Categórica
LONGITUD	Longitud geográfica de la institución	Numérica
LATITUD	Latitud geográfica de la institución	Numérica

Nota. Listado de las variables del conjunto de datos relativas a la ubicación del colegio

Población

Tabla 8

Variables de Población

Nombre de la Variable	Descripción	Tipo
Total	Total de la población del municipio	Numérica
Indígena	Población indígena	Numérica
Gitano(a) o Rrom	Población Gitana	Numérica
Raizal del Archipiélago de San Andrés, Providencia y Santa Catalina	Población raizal	Numérica
Palenquero(a) de San Basilio	Población Palenqueros	Numérica
Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)	Población Afro	Numérica
Ningún grupo étnico-racial	Población no perteneciente a ningún grupo étnico-racial	Numérica

Nota. Listado de las variables del conjunto de datos relativas a la población de los municipios

Herramientas y Tecnologías

Frontend

React: Una biblioteca de JavaScript ampliamente utilizada para construir interfaces de usuario dinámicas y reactivas. Facilita el desarrollo de componentes reutilizables y el manejo del estado de las aplicaciones.

TypeScript: Una extensión tipada de JavaScript que permite detectar errores antes de ejecutar el código, lo que mejora su robustez, escalabilidad y mantenibilidad.

shadcn/ui: Una biblioteca de componentes de interfaz de usuario que combina facilidad de uso con diseño moderno, simplificando la creación de interfaces intuitivas y consistentes.

Tailwind CSS: Un framework de CSS de utilidad que acelera el diseño de componentes, proporcionando clases predefinidas que permiten un estilizado rápido y personalizado.

Lucide React: Una colección de iconos vectoriales diseñados para React, utilizados para mejorar la experiencia visual y la claridad en las interfaces de usuario.

Backend

Flask: Un microframework de Python que permite desarrollar aplicaciones web y APIs de manera rápida y eficiente, ideal para proyectos ligeros o escalables.

Pandas: Biblioteca centrada en el análisis y manipulación de datos, ideal para gestionar estructuras tabulares de forma eficiente.

NumPy: Herramienta para operacionesNuméricas avanzadas, esencial para manejar datos multidimensionales.

Scipy: Una biblioteca robusta para cálculos estadísticos y matemáticos avanzados, usada para realizar pruebas y análisis estadísticos.

Statsmodels: Incluye funcionalidades avanzadas para análisis estadísticos, como pruebas de normalidad y gráficos de diagnóstico.

Matplotlib: Una herramienta versátil para generar visualizaciones personalizadas de datos.

Seaborn: Complemento de Matplotlib que facilita la creación de gráficos estadísticos estilizados y efectivos.

Scikit-learn: Una biblioteca para machine learning, utilizada aquí para tareas como regresión lineal.

Folium: Biblioteca para generar mapas interactivos, combinada con plugins como MarkerCluster para agregar funcionalidad avanzada.

Google Generative AI: Herramienta para integraciones avanzadas con inteligencia artificial generativa, ideal para interpretar resultados y generar insights.

Chardet: Utilizada para detectar la codificación de archivos de texto, asegurando la correcta lectura de datos.

Pathlib: Simplifica la manipulación de rutas de archivos en sistemas operativos.

Base64: Permite codificar y decodificar datos binarios para su transferencia o almacenamiento.

IO y BytesIO: Facilitan el manejo de flujos de entrada y salida, como archivos en memoria.

Base de Datos

Archivos CSV: Los datos serán gestionados a través de archivos en formato CSV, un estándar ampliamente utilizado para almacenar y transferir información estructurada de forma tabular.

Herramientas de Desarrollo

Git: Un sistema de control de versiones distribuido que permite gestionar cambios en el código de forma eficiente y colaborar con equipos de desarrollo.

npm: El gestor de paquetes oficial para el ecosistema JavaScript, utilizado para instalar y administrar dependencias en proyectos basados en React o TypeScript.

pip: El gestor de paquetes de Python que facilita la instalación de bibliotecas necesarias para el backend, como Flask, Pandas o NumPy.

Visual Studio Code: Un editor de código ligero y multiplataforma que soporta una amplia gama de lenguajes y herramientas, ideal para desarrollo frontend y backend.

Metodología de Desarrollo

En este proyecto se utilizó el marco de trabajo Scrum para garantizar una gestión ágil y efectiva del desarrollo, asegurando que los objetivos se lograran de manera iterativa e incremental. El proyecto se dividió en sprints, cada uno enfocado en implementar un caso de uso específico. Durante cada sprint, se desarrollaron de manera simultánea las funcionalidades requeridas en el frontend, la API y el backend, permitiendo entregar incrementos funcionales al final de cada iteración.

Se adoptaron las ceremonias clásicas de Scrum, comenzando con una planificación del sprint en la que se definieron los requisitos del caso de uso a implementar, como la segmentación de datos o la generación de visualizaciones. En el transcurso del sprint, se realizaron revisiones diarias para alinear avances y resolver bloqueos, mientras que, al cierre, las revisiones del sprint permitieron evaluar los resultados y recopilar retroalimentación. Esta metodología facilitó la adaptación a nuevos requerimientos y la mejora continua. A continuación, se describe cada uno de los casos de uso o funcionalidades del sistema y se presentan los diagramas básicos de cada uno. Al final se presentan los diagramas de componentes y clases y su interacción.

Casos de Uso

A continuación, se describen los casos de uso del sistema, presentando para cada uno las capturas de pantalla y los diagramas principales.

Particionar Conjuntos de Datos

Esta función permite a los usuarios dividir o filtrar conjuntos de datos más grandes en subconjuntos más manejables. Los usuarios pueden acceder a la funcionalidad de particionamiento de datos a través de la pestaña "Particionar" que se muestra en la Figura 1.

Figura 1

Particionamiento de Conjunto de Datos

Nota. Pantalla para particionamiento del set de datos colocando filtros por fila.

El usuario selecciona un conjunto de datos, selecciona los filtros que desee aplicar al conjunto de datos inicial que consta del universo de resultados de los años 2014 a 2023 de las pruebas ICFES Saber 11. Puede seleccionar otros conjuntos de datos que haya generado en la misma sesión de trabajo. Una vez que ha establecido los filtros, puede dar click al botón “Aplicar Filtros” para visualizar el resultado en la pantalla. La **Figura 4** presenta el diagrama de casos de uso para este componente donde lo podemos ver gráficamente. La Figura 5 muestra el diagrama de secuencia de este caso de uso con los nombres de los componentes del código fuente para una mejor comprensión.

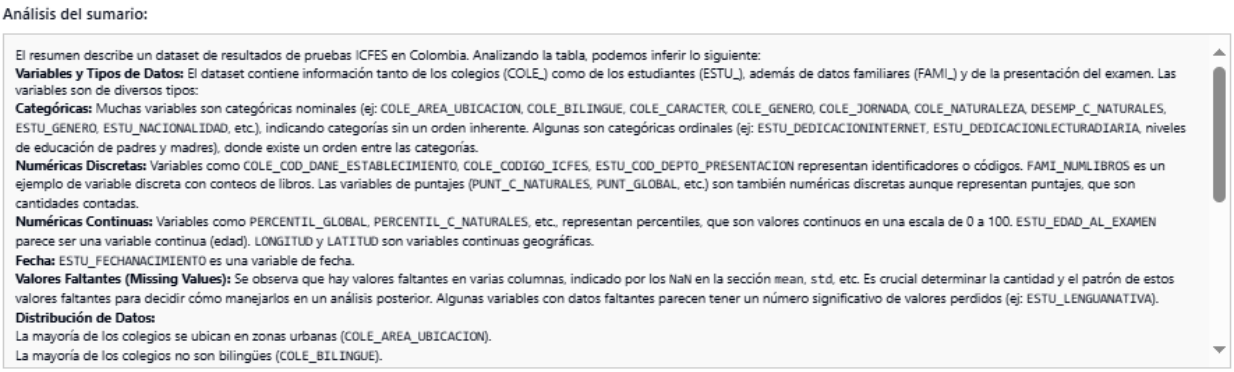
En la se observa un ejemplo donde se ha seleccionado el conjunto de datos llamado “Bogotá – 2033.csv” y se le ha aplicado el filtro COLE_CALENDARIO = ‘A’. En este momento

el usuario tiene la opción de ingresar el nombre del set de datos ya filtrado y salvarlo en el servidor con el botón “Guardar Dataset.

Si el usuario desea, puede hacer click en el botón “Analizar Resumen”, donde el sistema realiza un análisis asistido por inteligencia artificial. La Figura 2 presenta el resultado del análisis con algunas recomendaciones.

Figura 2

Resultados del Análisis de Sumario

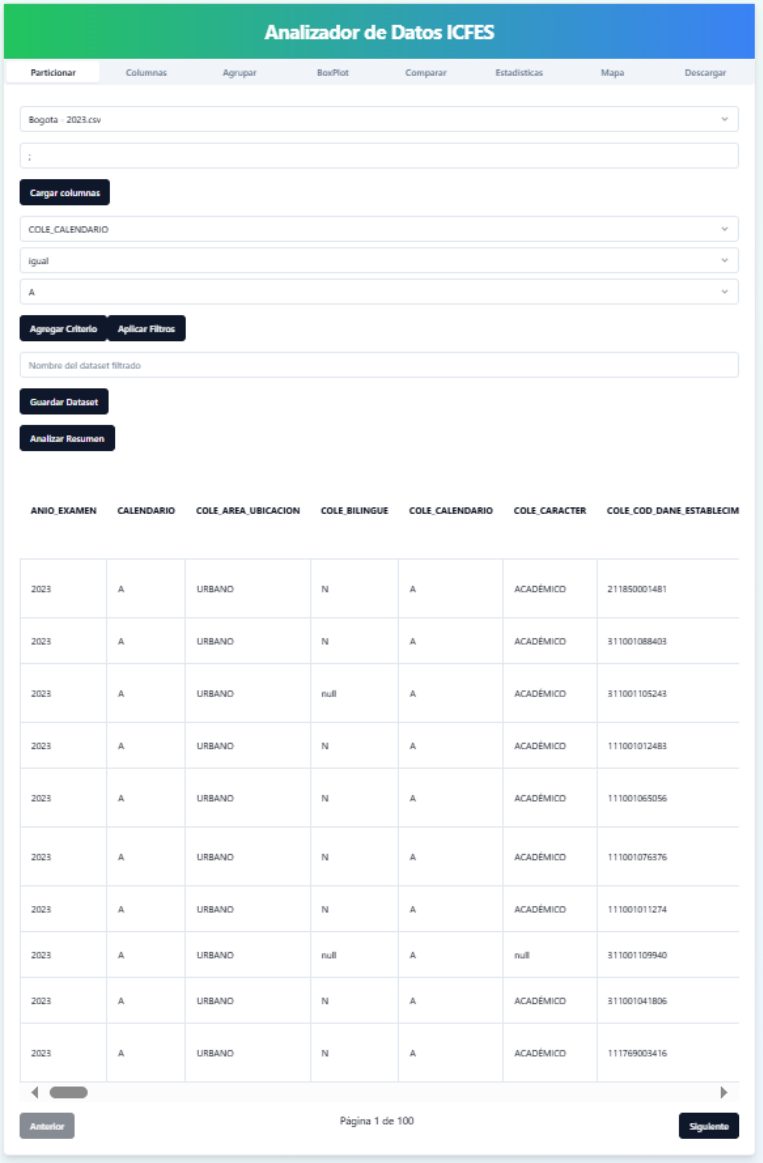


Nota. Resultado del análisis del sumario del conjunto de datos asistido por IA

La Figura 4 presenta el diagrama de casos de uso para este componente donde lo podemos ver gráficamente. La Figura 5 muestra el diagrama de secuencia de este caso de uso con los nombres de los componentes del código fuente para una mejor comprensión.

Figura 3

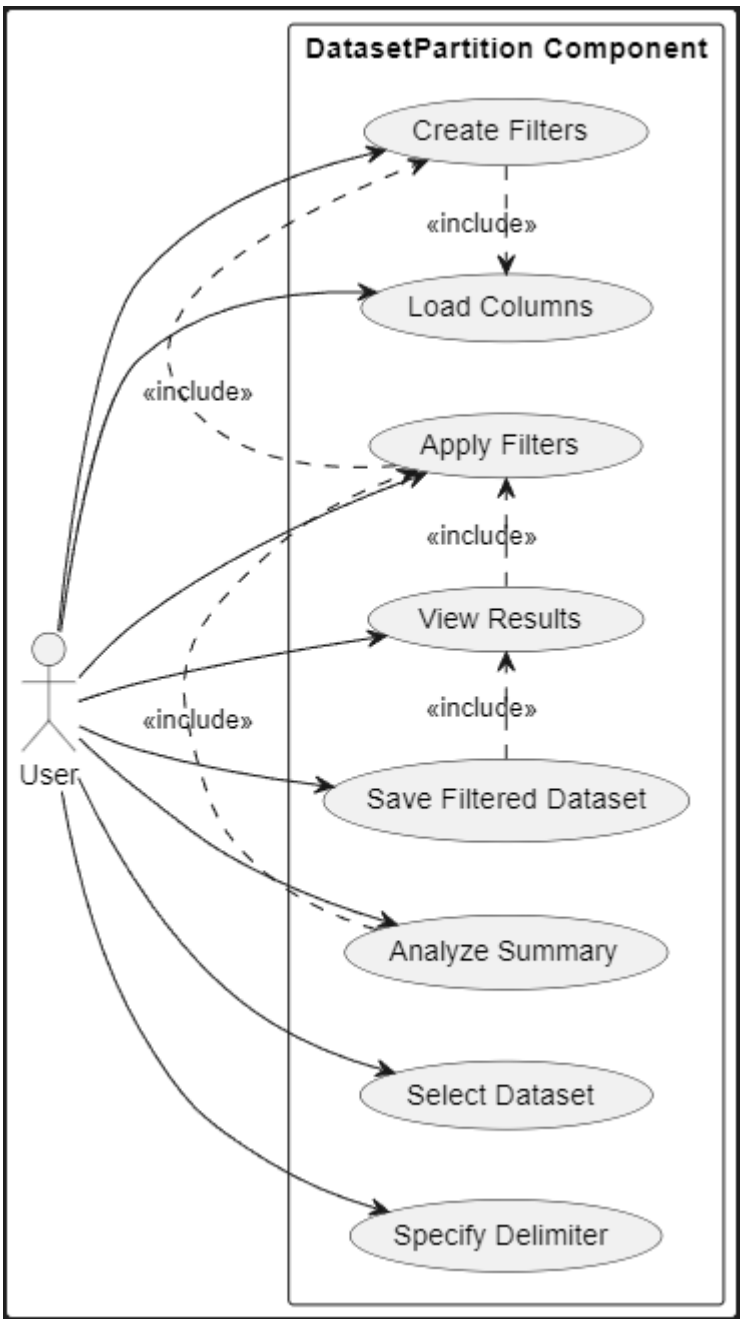
Aplicar Filtros



Nota. Opciones para aplicar filtros dando condiciones a una o varias columnas

Figura 4

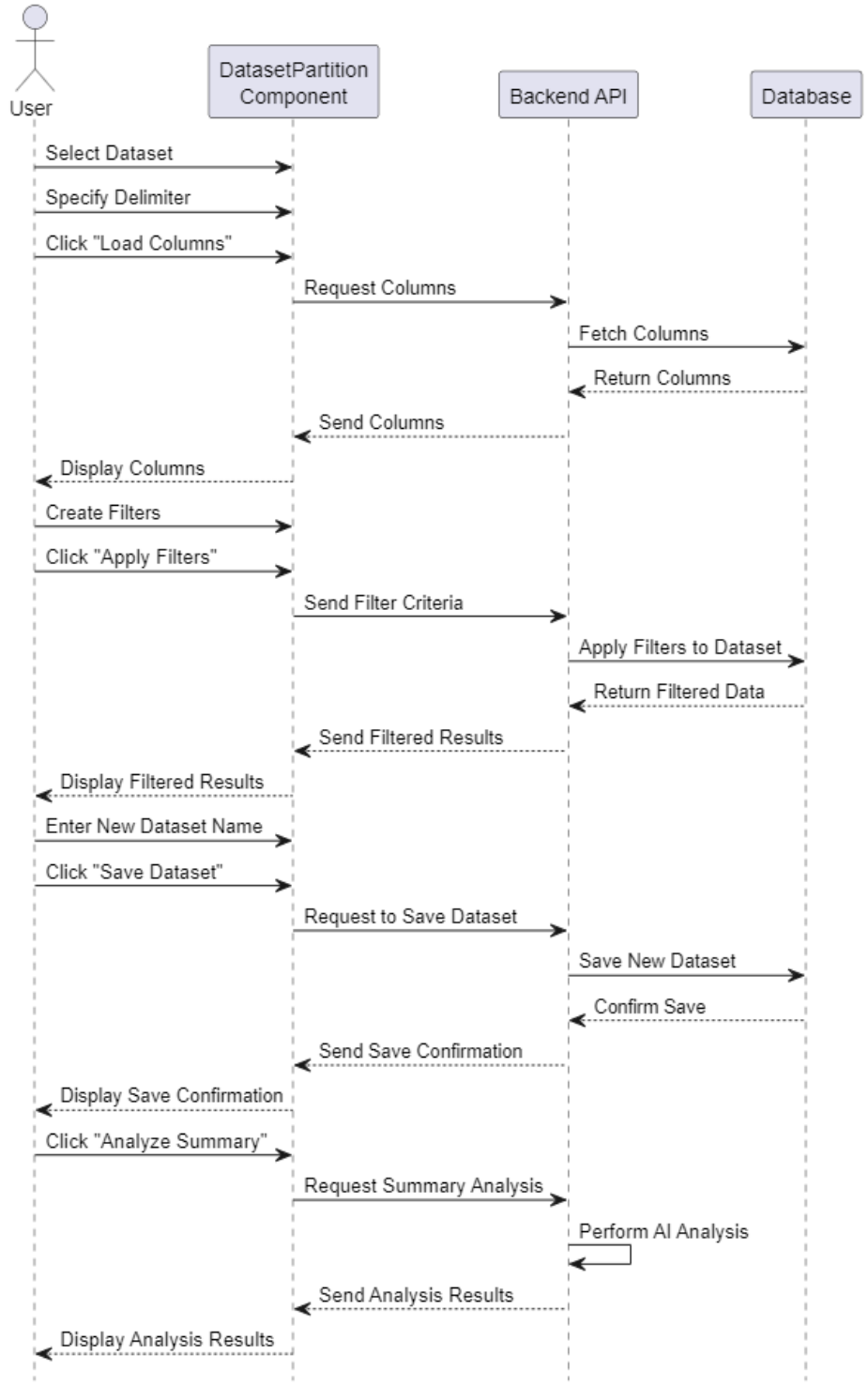
Caso de Uso - Particionar Datos



Nota. Diagrama de casos de uso para particionar datos aplicando filtros

Figura 5

Diagrama de Secuencia – Particionar Datos



Nota. Diagrama de secuencia de la opción de particionar datos aplicando filtros

Seleccionar Columnas

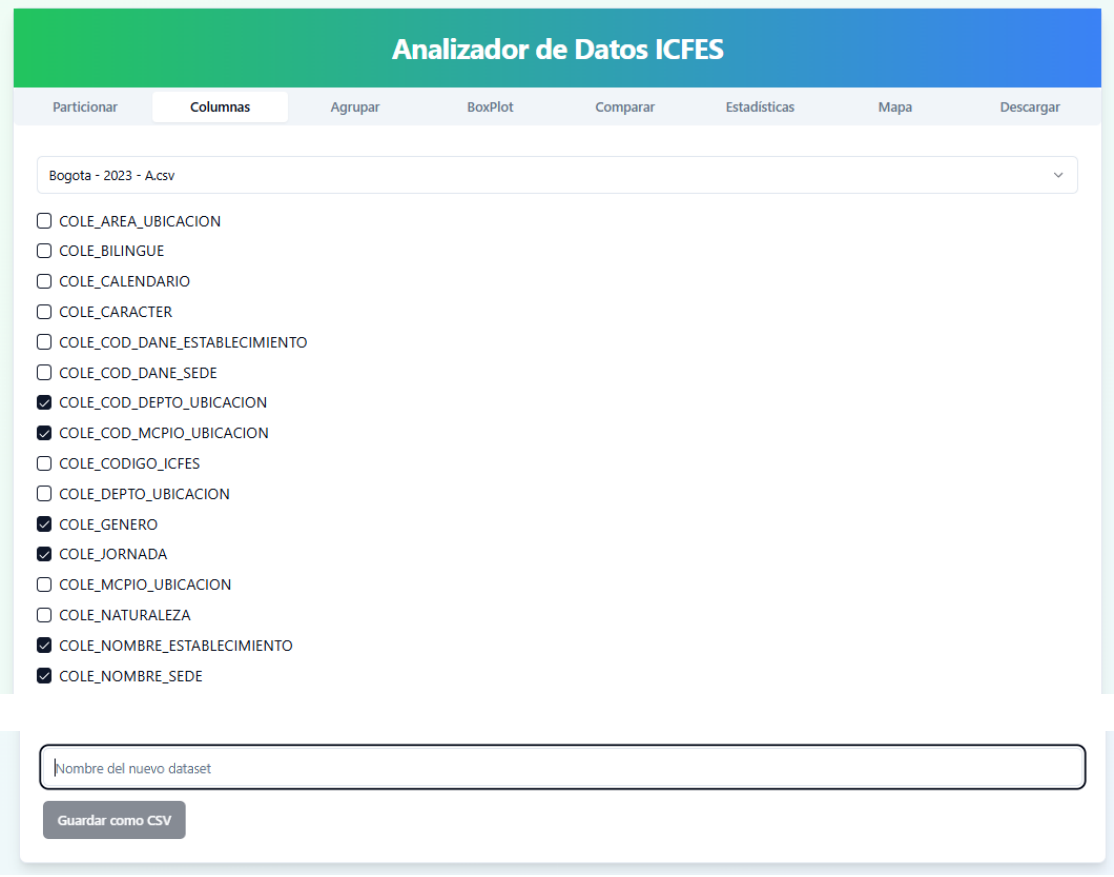
Los usuarios pueden seleccionar columnas específicas de los conjuntos de datos para su análisis o visualización. Accesible mediante la pestaña "Columnas"

Al igual que en la opción anterior, el usuario puede seleccionar un conjunto de datos y luego seleccionar las columnas que necesita en su archivo final.

En la Figura 6 se presenta la selección de algunas columnas del conjunto de datos. En la parte inferior de la pantalla aparece la opción para nombrar el nuevo conjunto de datos y salvarlo.

Figura 6

Selector de Columnas

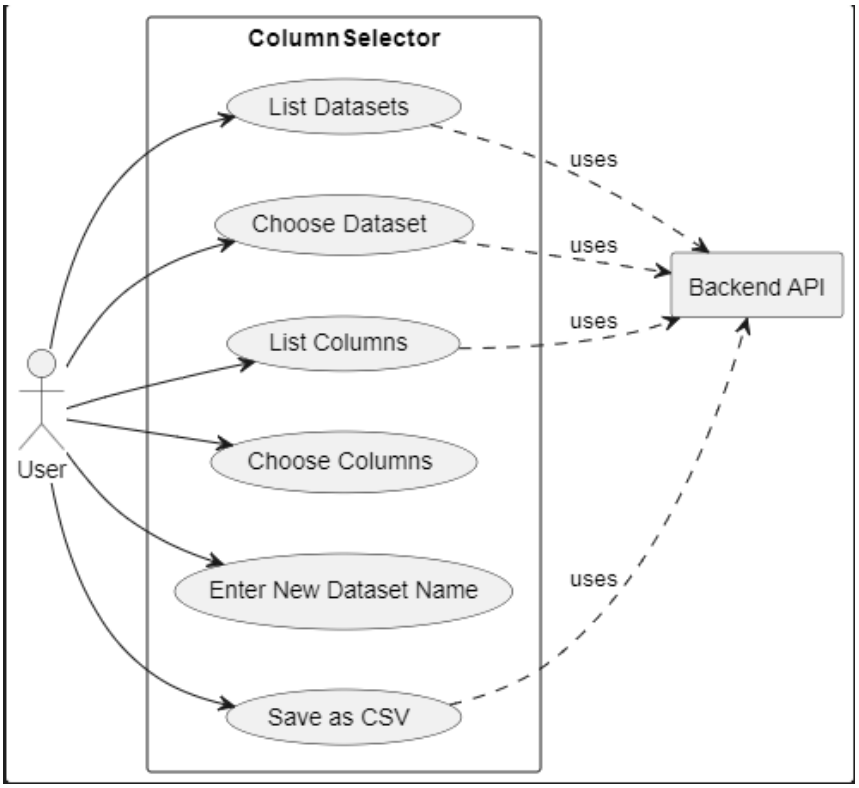


Nota. Opciones para seleccionar las columnas para generar un nuevo conjunto de datos reducido

A continuación, se presentan los diagramas de casos de uso y secuencia de este componente.

Figura 7

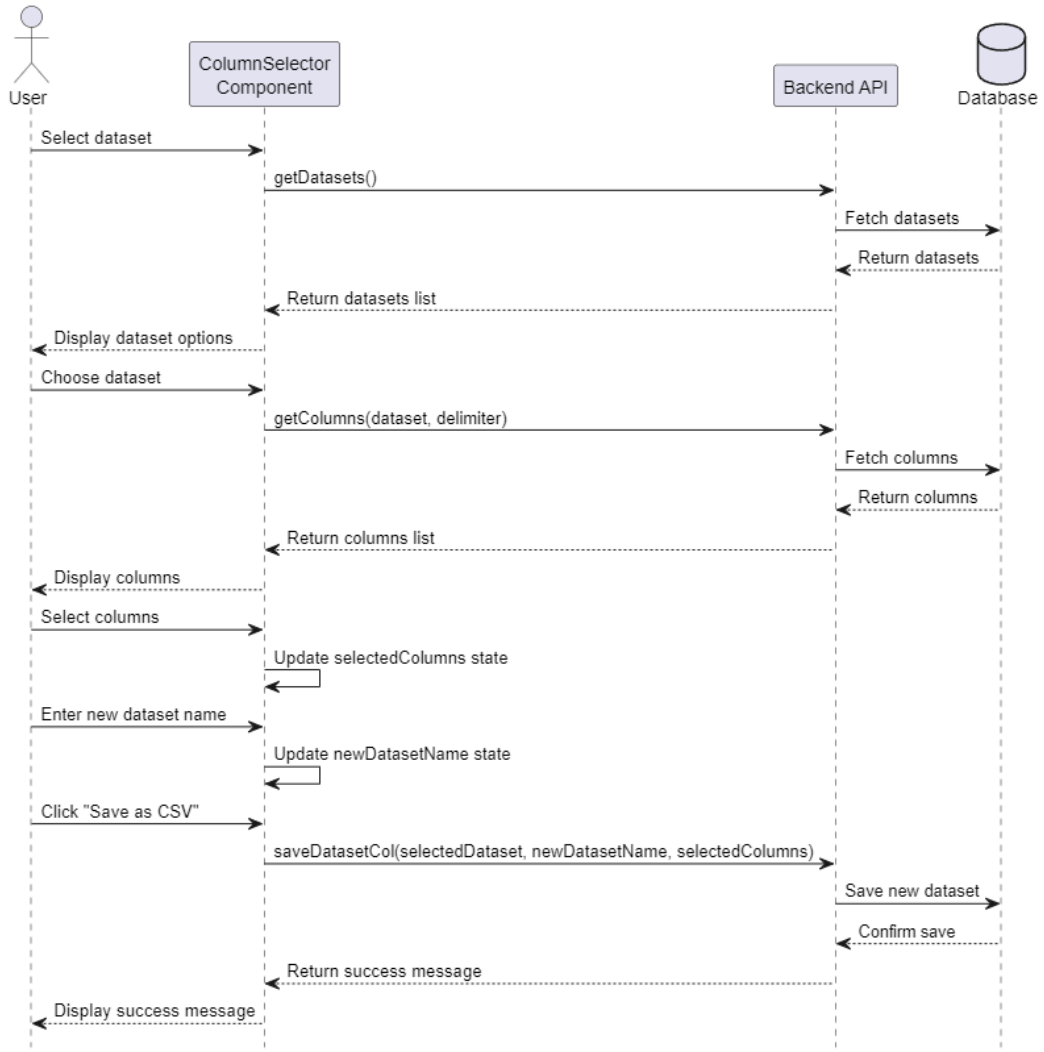
Caso de Uso - Seleccionar Columnas



Nota. Diagrama con las opciones generar un conjunto de datos a partir de las columnas seleccionadas por el usuario.

Figura 8

Diagrama de Secuencia - Seleccionar Columnas



Nota. Diagrama de la secuencia de eventos para realizar el filtro por columnas de un conjunto de datos y salvarlo

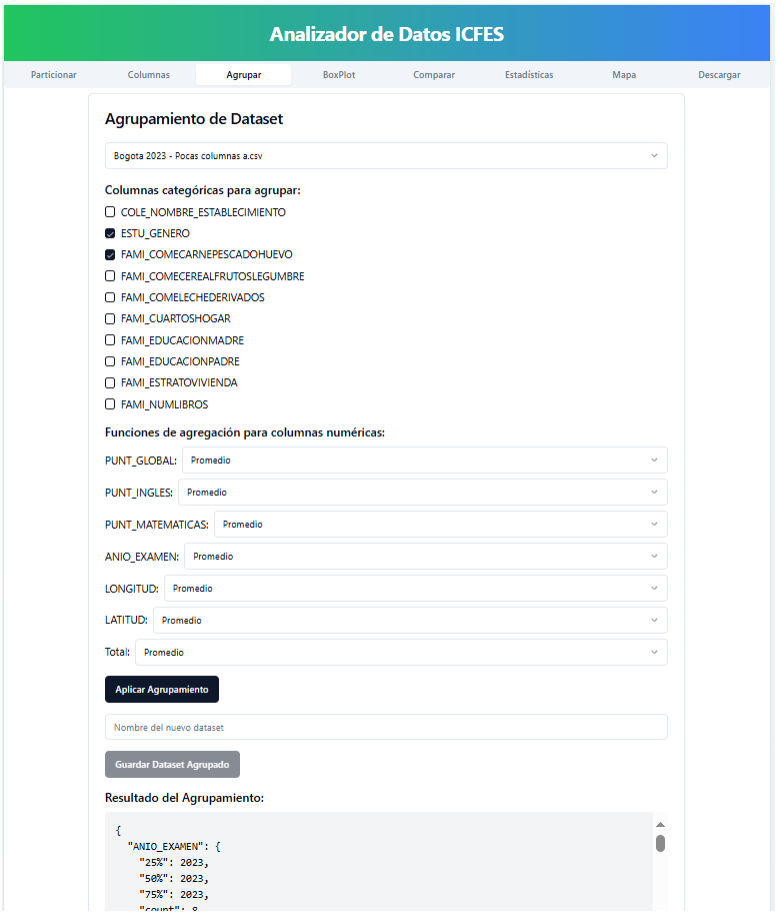
Agrupar Datos

Se accede a través de la pestaña "Agrupar". Permite a los usuarios agrupar datos basándose en criterios específicos, facilitando el análisis de tendencias o patrones en subgrupos de datos. La funcionalidad consiste en seleccionar un conjunto de datos y luego seleccionar las

columnas para realizar el agrupamiento y, en segundo lugar, seleccionar las funciones agregadas que se aplicarán a las columnas Numéricas. En la Figura 9 se presenta la pantalla donde se toma el conjunto de datos que se generó en el ejercicio del apartado anterior y se le aplica un agrupamiento por las variables de género y la variable que indica la frecuencia con la que comen carne, pescado o huevos en la familia. Al aplicar el agrupamiento se obtiene en la parte inferior en formato JSON las estadísticas básicas de cada una de las variables del set de datos y se ofrece la opción de guardar este nuevo resultado.

Figura 9

Pantalla de Agrupamiento de Datos

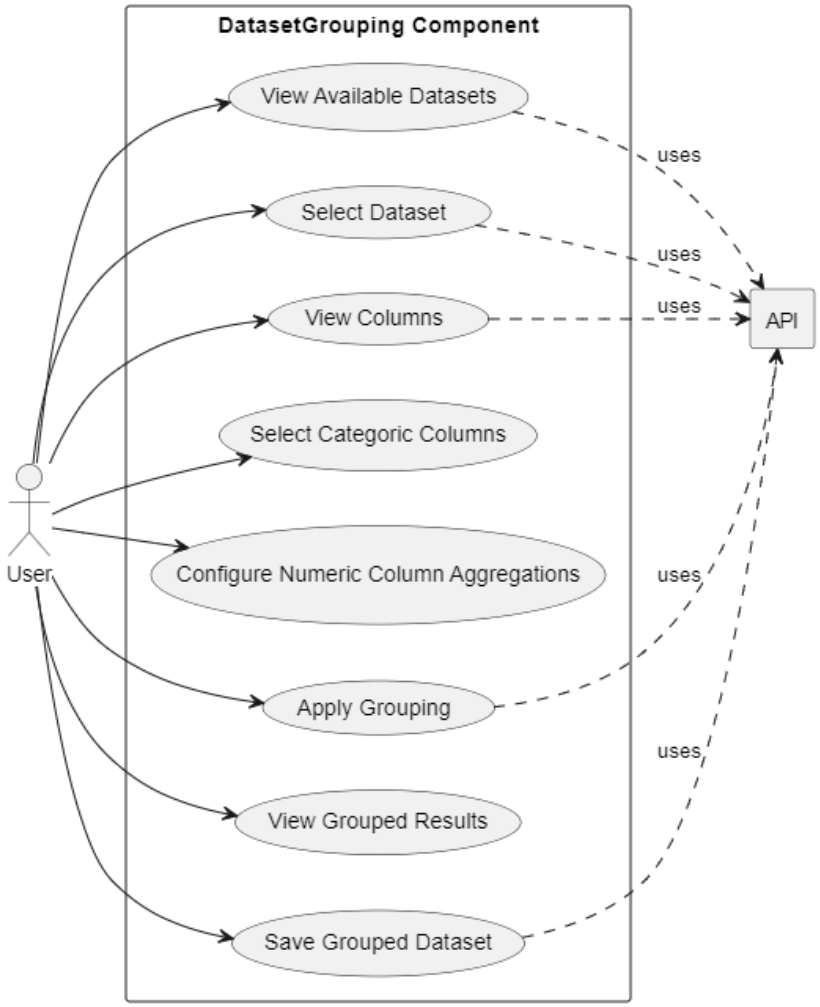


Nota. Pantalla con las opciones de selección de columnas para agrupar y funciones agregadas a aplicar

A continuación, se presentan los diagramas de casos de uso y secuencia para el componente.

Figura 10

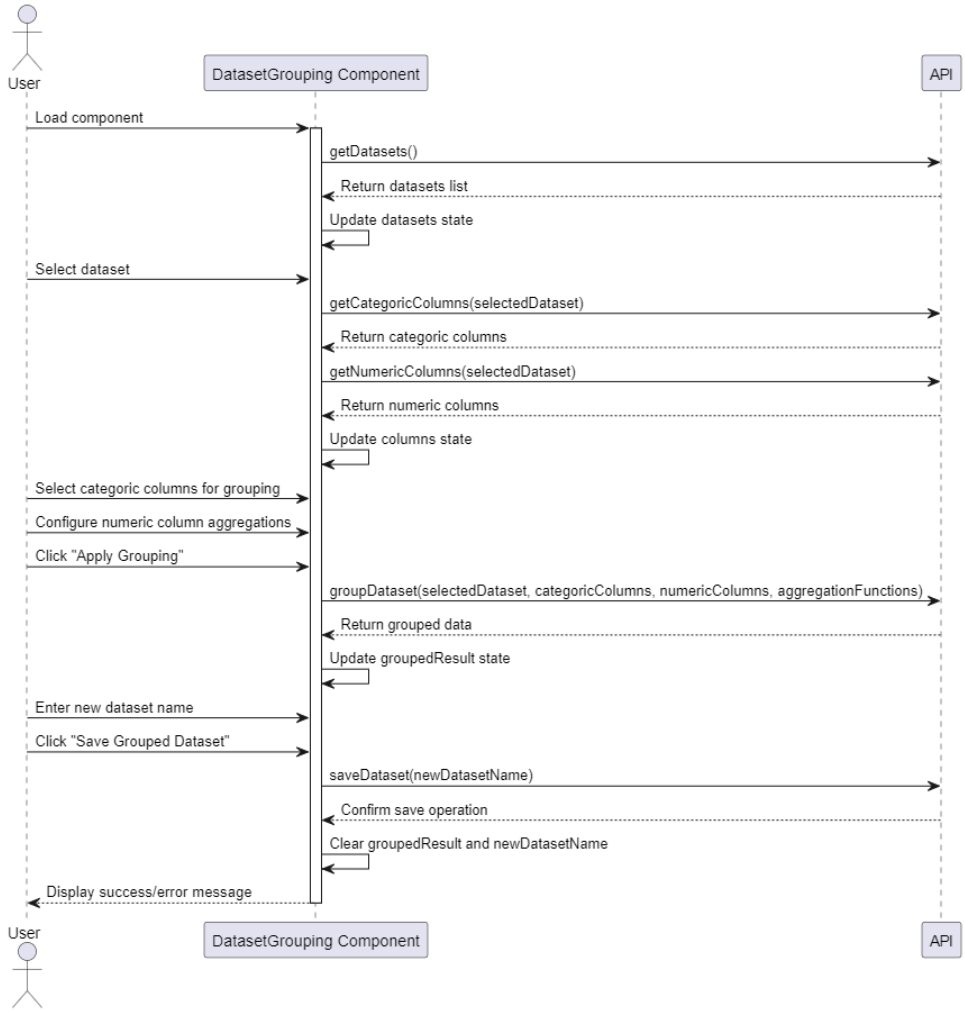
Caso de uso – Agrupar



Nota. Diagrama de uso de la función de realizar agrupamientos por las columnas seleccionadas por el usuario

Figura 11

Diagrama de Secuencia – Agrupar



Nota. Diagrama con la secuencia de eventos del sistema cuando se aplica la opción de agrupar el conjunto de datos por unas columnas seleccionadas

Generar Box Plots

Disponible en la pestaña "BoxPlot". Los usuarios pueden crear diagramas de caja y bigotes para visualizar la distribución de datos e identificar outliers. La funcionalidad consiste en seleccionar un conjunto de datos, luego una variable Categórica y una variable Numérica. En la

Figura 12 se presenta un boxplot de la variable COLE_GENERO contra la variable PUNT_GLOBAL.

La aplicación presenta un botón con el que se puede solicitar un análisis del boxplot asistido por inteligencia artificial.

Figura 12

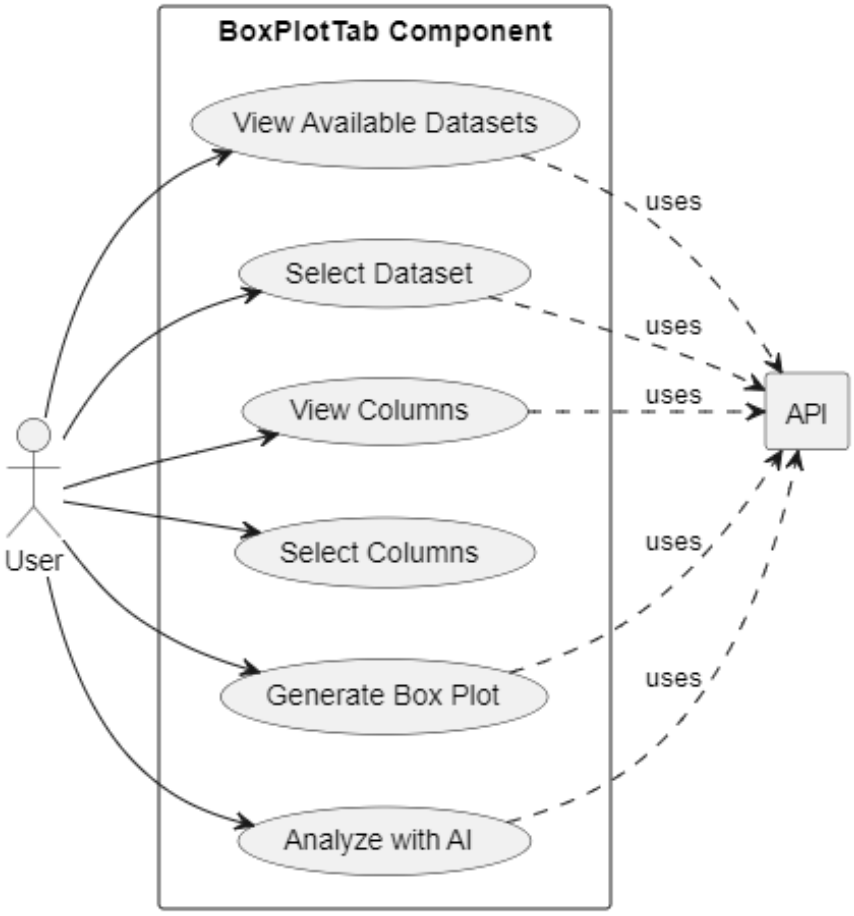
Generación de Box Plot

Nota. Pantalla del sistema que permite generar un box plot seleccionando una variable Categórica y una Numérica

En el ejemplo de la Figura 12, podemos ver que la media del puntaje global de los colegios mixtos en menor que los demás, lo cual coincide con el análisis asistido por inteligencia artificial.

Figura 13

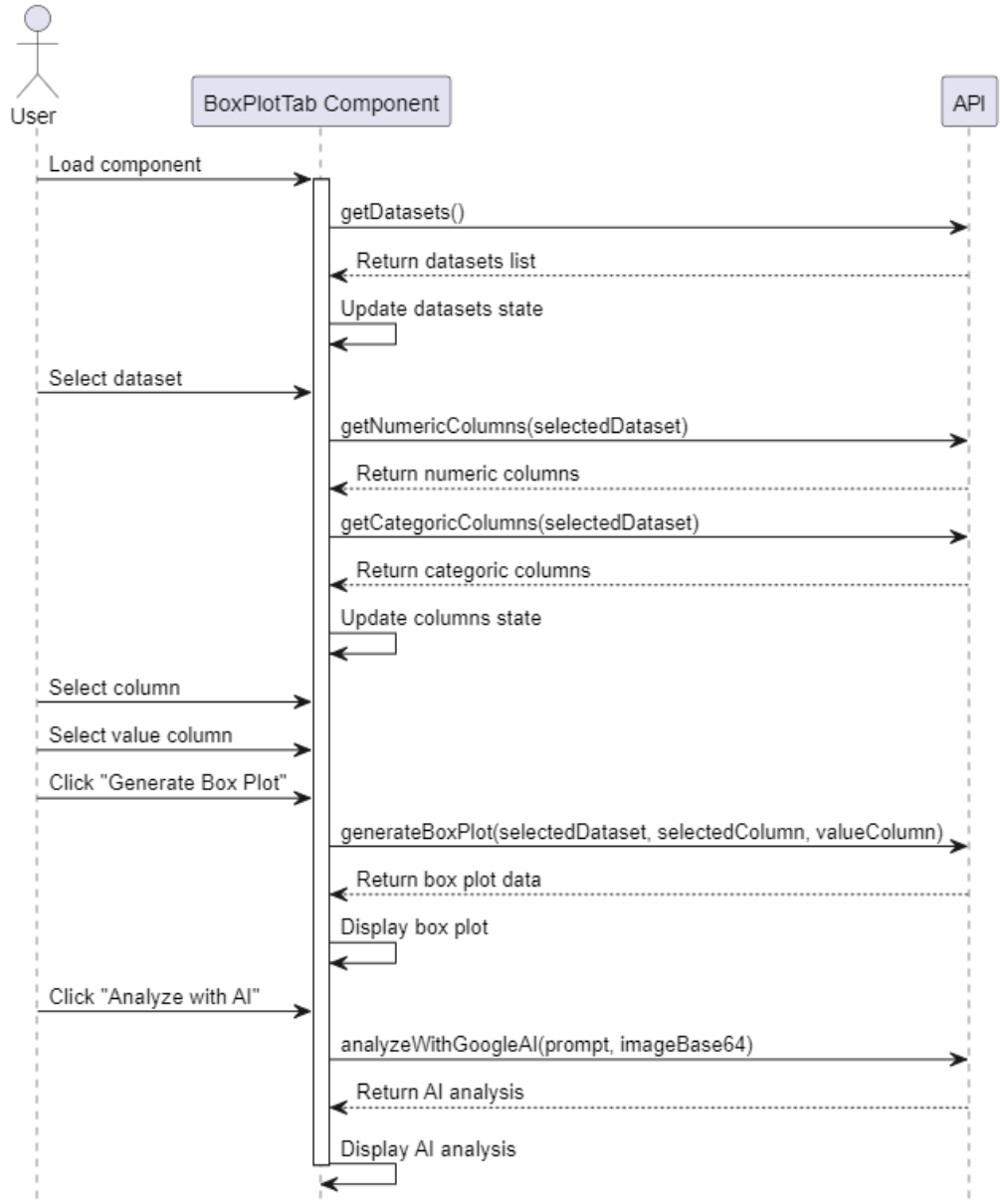
Caso de Uso - Generar Box Plot



Nota. Diagrama de uso de la opción para generar un box plot sobre las columnas seleccionada

Figura 14

Diagrama de Secuencia - Generar Box Plot



Nota. Diagrama de los eventos generados por la opción de generar y analizar box plots.

Generar Análisis Comparativo

Se accede mediante la pestaña "Comparar". Permite a los usuarios comparar diferentes conjuntos de datos o variables, a través de gráficos. En la Figura 15 se observa un comparativo de los conjuntos de datos de Nariño, comparado las frecuencias con los datos de Bogotá, calendario A. Las variables para comparar pueden ser Numéricas o Categóricas. Cuando se seleccionan las variables Numéricas, el gráfico presenta una distribución de frecuencias.

Figura 15

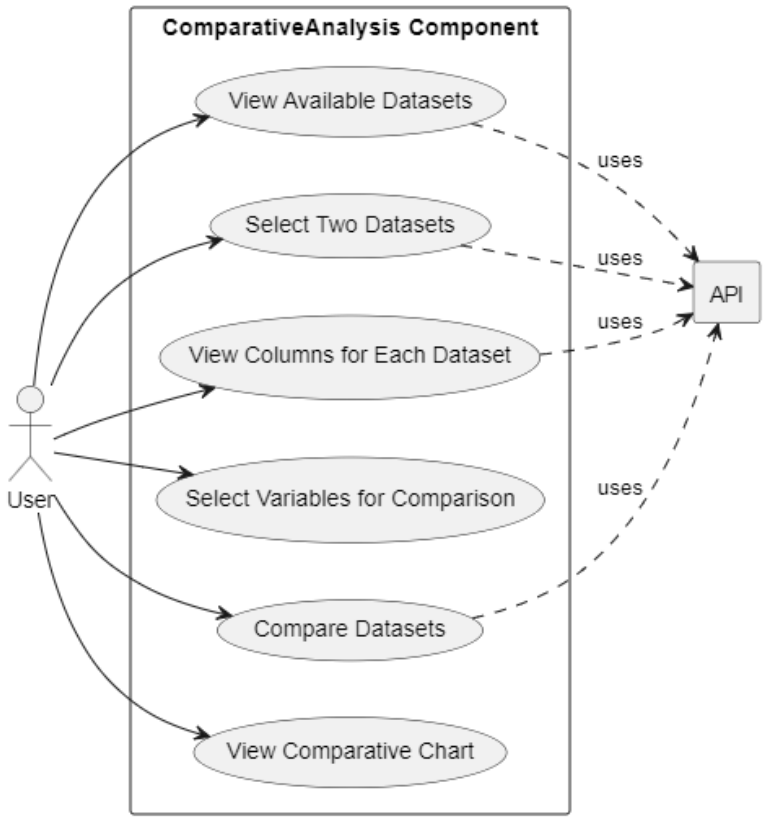
Elaborar Análisis Comparativo

Nota. Pantalla del sistema para realizar un comparativo entre variables de diferente conjunto de datos.

Podemos ver en el ejemplo las proporciones de hombres y mujeres entre los dos conjuntos de datos seleccionados. En la Figura 16 y en la Figura 17 se observan los diagramas de caso de uso y secuencia, respectivamente, de la funcionalidad de este componente.

Figura 16

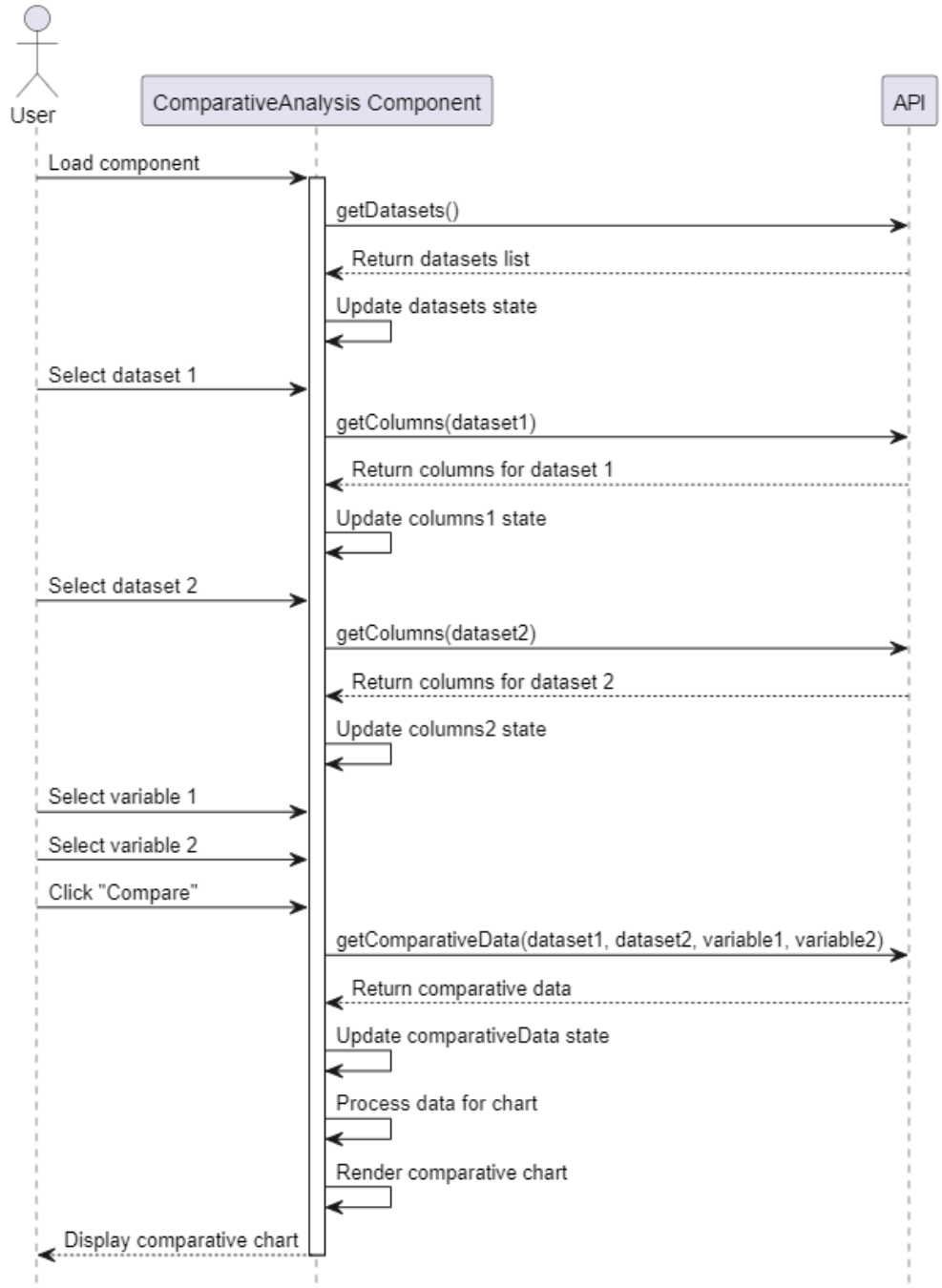
Caso de Uso - Generar Análisis Comparativo



Nota. Diagrama de uso de la opción para generar gráficos comparativos de variables de diferente conjunto de datos.

Figura 17

Diagrama de Secuencia - Generar Comparativo

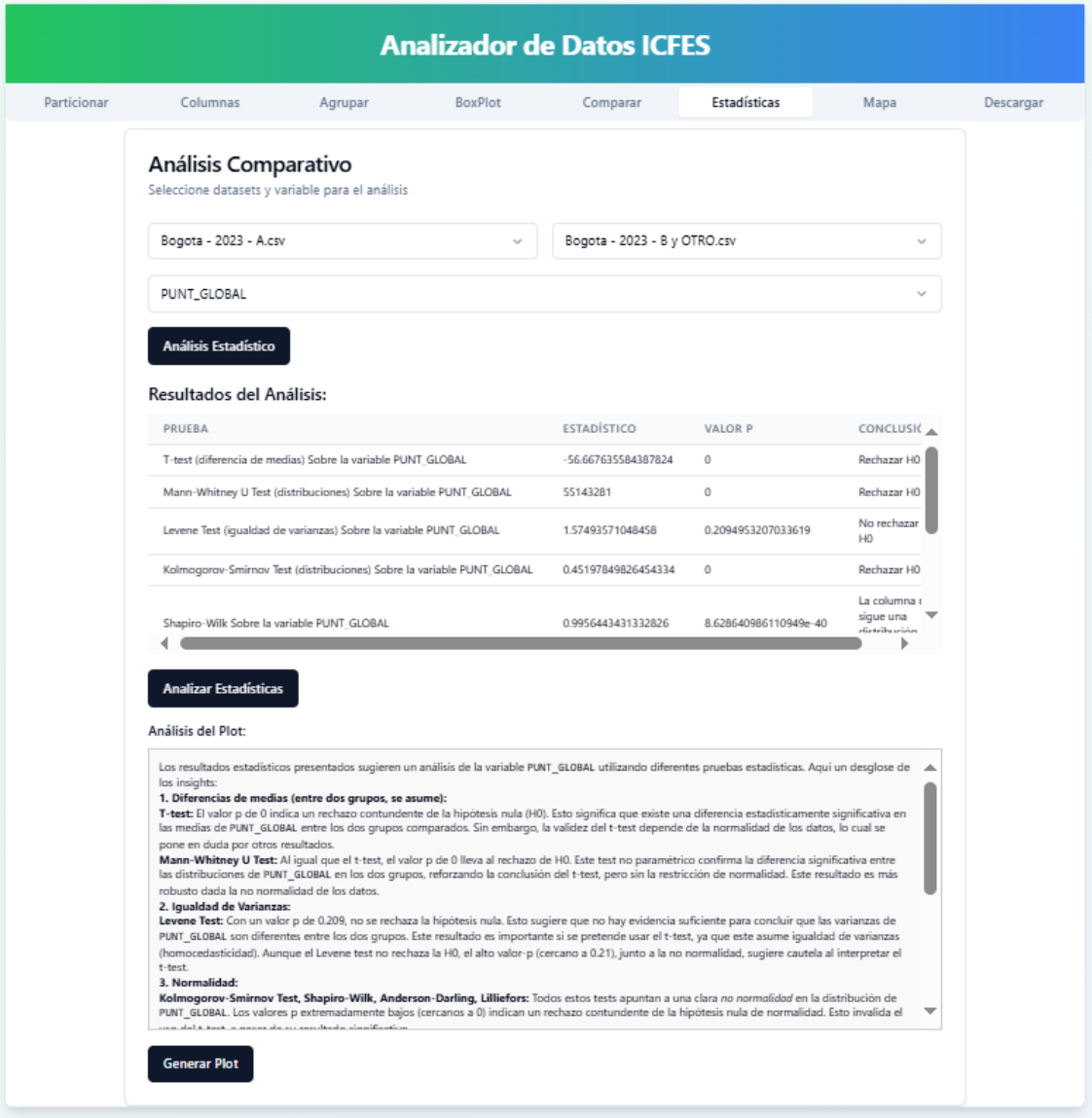


Nota. Diagrama de eventos de la opción para generar las gráficas comparativas

Generar Estadísticas Comparativas

Figura 18

Generación de Estadísticas Comparativas



Nota. Pantalla con las opciones de generación de análisis estadísticos comparativos entre diferentes conjuntos de datos

Accesible a través de la pestaña "Estadísticas". Ofrece análisis estadísticos comparativos, mostrando métricas clave para diferentes grupos o variables. El usuario debe seleccionar dos conjuntos de datos para comparar y una variable de referencia. Al oprimir el botón Análisis estadístico, se genera una tabla donde se aplican las principales pruebas estadísticas tanto para la variable como para el comparativo. En la Figura 18 se observa la tabla de resultado y sus correspondientes análisis.

Análisis Comparativo

Evaluar diferencias entre dos columnas de diferentes conjuntos de datos, dependiendo de si los datos son numéricos o categóricos.

Pruebas aplicadas:

- T-Test (Prueba de diferencia de medias):

Propósito: Comparar las medias de dos grupos independientes.

Hipótesis nula (H0): No hay diferencia significativa entre las medias.

Rechazo de H0: Si pp < 0.05.

- Mann-Whitney U:

Propósito: Comparar distribuciones de dos grupos cuando no se puede asumir normalidad.

H0: Las distribuciones de ambos grupos son iguales.

Rechazo de H0: Si pp < 0.05.

- Prueba de Levene:

Propósito: Verificar la igualdad de varianzas entre dos grupos.

H0: Las varianzas son iguales.

Rechazo de H0: Si pp < 0.05.

- Kolmogorov-Smirnov (KS):

Propósito: Comparar si dos distribuciones son iguales.

H0: No hay diferencia entre las distribuciones.

Rechazo de H0: Si pp < 0.05.

- Chi-cuadrado (para datos categóricos):

Propósito: Evaluar la independencia entre dos variablesCategóricas.

H0: Las variables son independientes.

Rechazo de H0: Si pp < 0.05.

- Prueba Exacta de Fisher (para datos categóricos con tamaños pequeños):

Propósito: Evaluar independencia entre variablesCategóricas.

H0: Las variables son independientes.

Rechazo de H0: Si pp < 0.05.

Análisis de una Columna

Analizar características estadísticas de una única columna en un conjunto de datos.

Pruebas aplicadas:

- Shapiro-Wilk:

Propósito: Verificar normalidad de los datos.

H0: Los datos siguen una distribución normal.

Rechazo de H0: Si pp < 0.05.

- Anderson-Darling:

Propósito: Evaluar si los datos siguen una distribución específica (normal en este caso).

H0: Los datos son consistentes con una distribución normal.

Rechazo de H0: Si el estadístico supera el valor crítico.

- Kolmogorov-Smirnov (KS):

Propósito: Comparar si los datos siguen una distribución normal.

H0: La distribución es normal.

Rechazo de H0: Si pp < 0.05.

- Lilliefors:

Propósito: Verificar normalidad, ajustada para parámetros desconocidos.

H0: Los datos son normales.

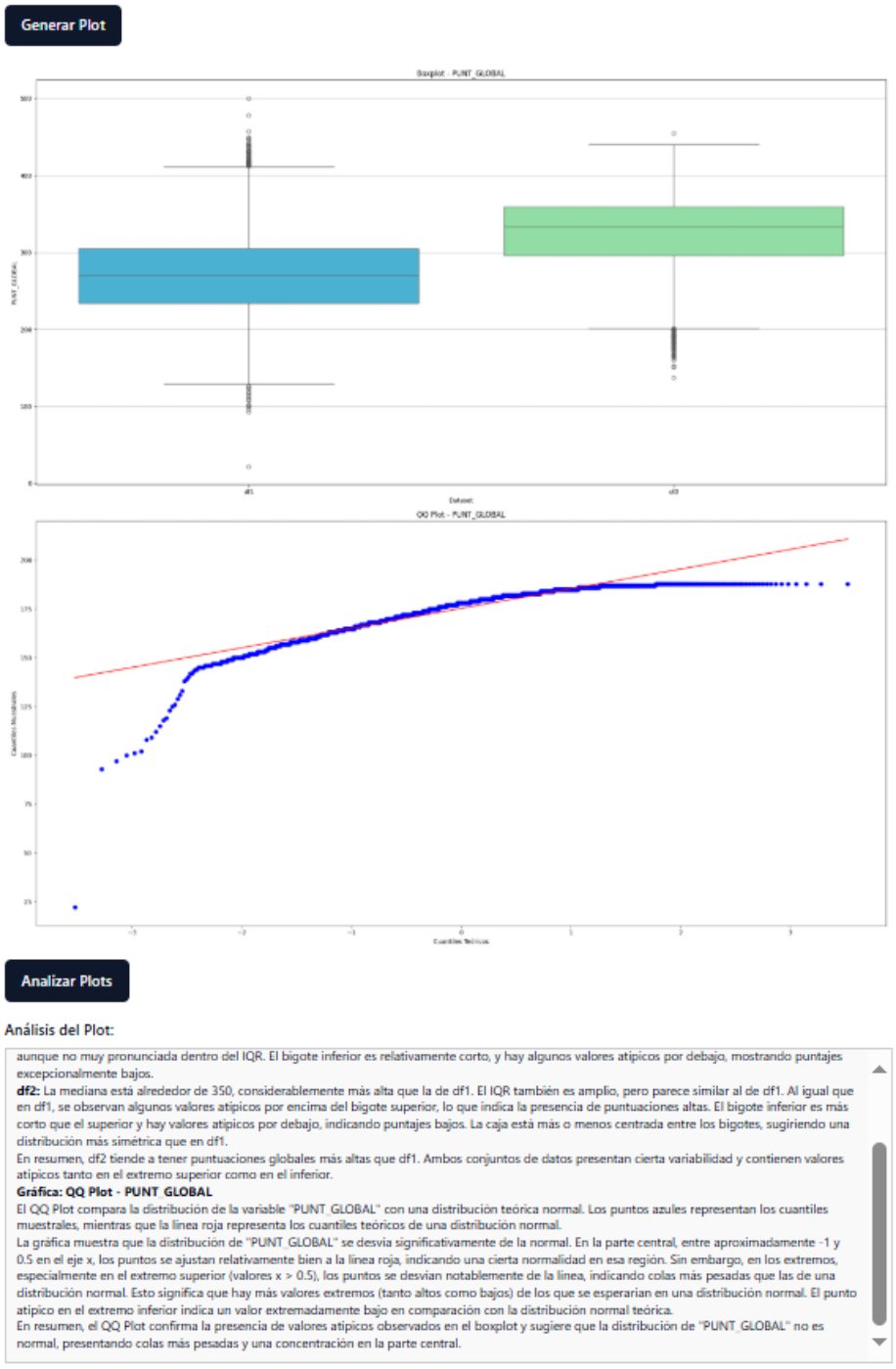
Rechazo de H0: Si pp < 0.05.

- Distribución de categorías (para datos categóricos):

Se calculan las frecuencias de cada categoría y se presenta su distribución. Cuando el usuario ha generado la tabla de pruebas estadísticas, se puede generar el análisis asistido por inteligencia artificial o generar una gráfica comparativa. En la Figura 19 vemos el Box Plot comparativo de la variable PUNT_GLOBAL para los dos sets de datos seleccionados y en la parte inferior vemos un QQ-Plot. El botón para analizar los plot entrega una explicación de las imágenes generadas.

Figura 19

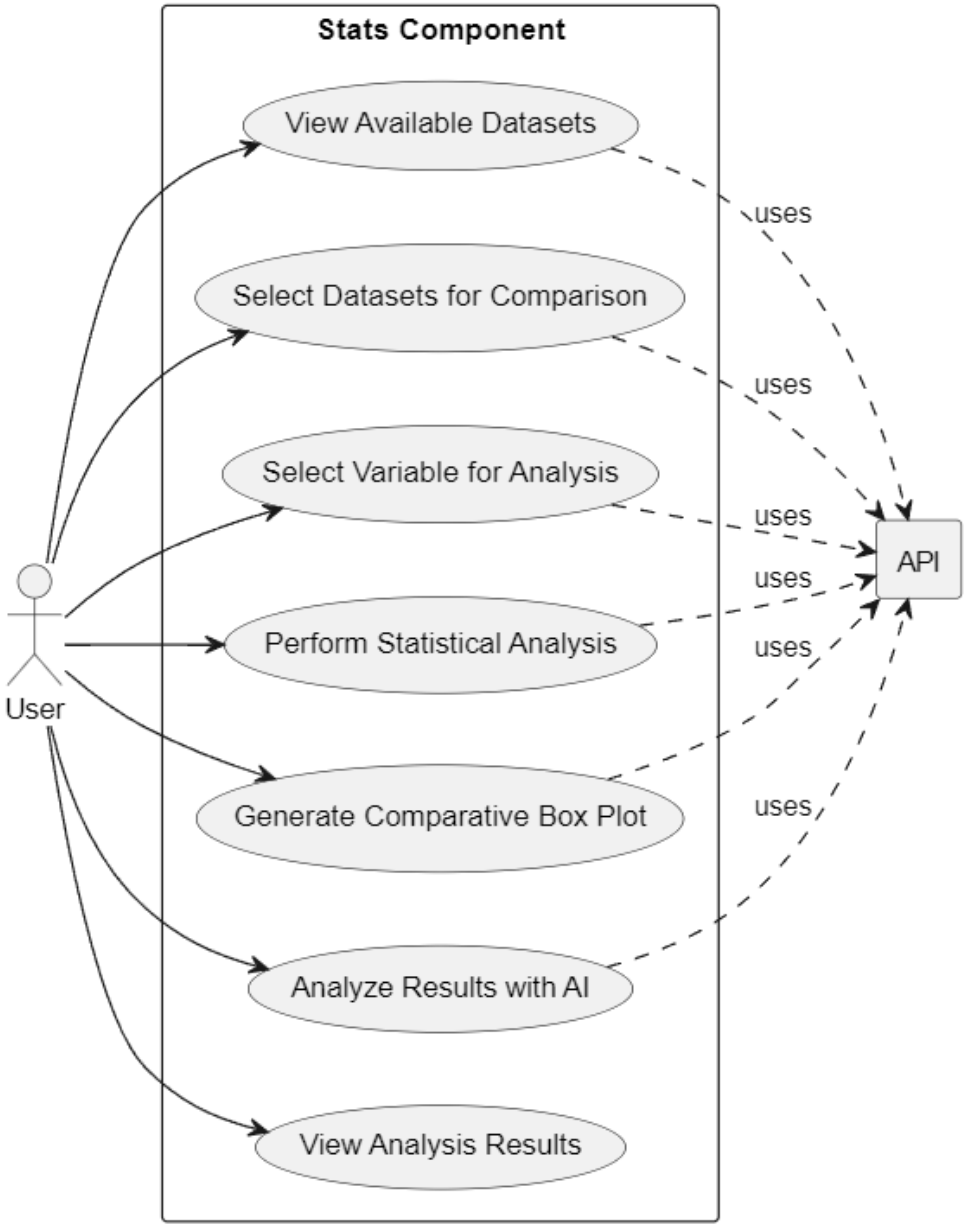
Plot Comparativo Generado



Nota. Pantalla del box plot comparativo y el qq-plot generados a partir de la variable seleccionada.

Figura 20

Caso de uso - Generar Estadísticas Comparativas

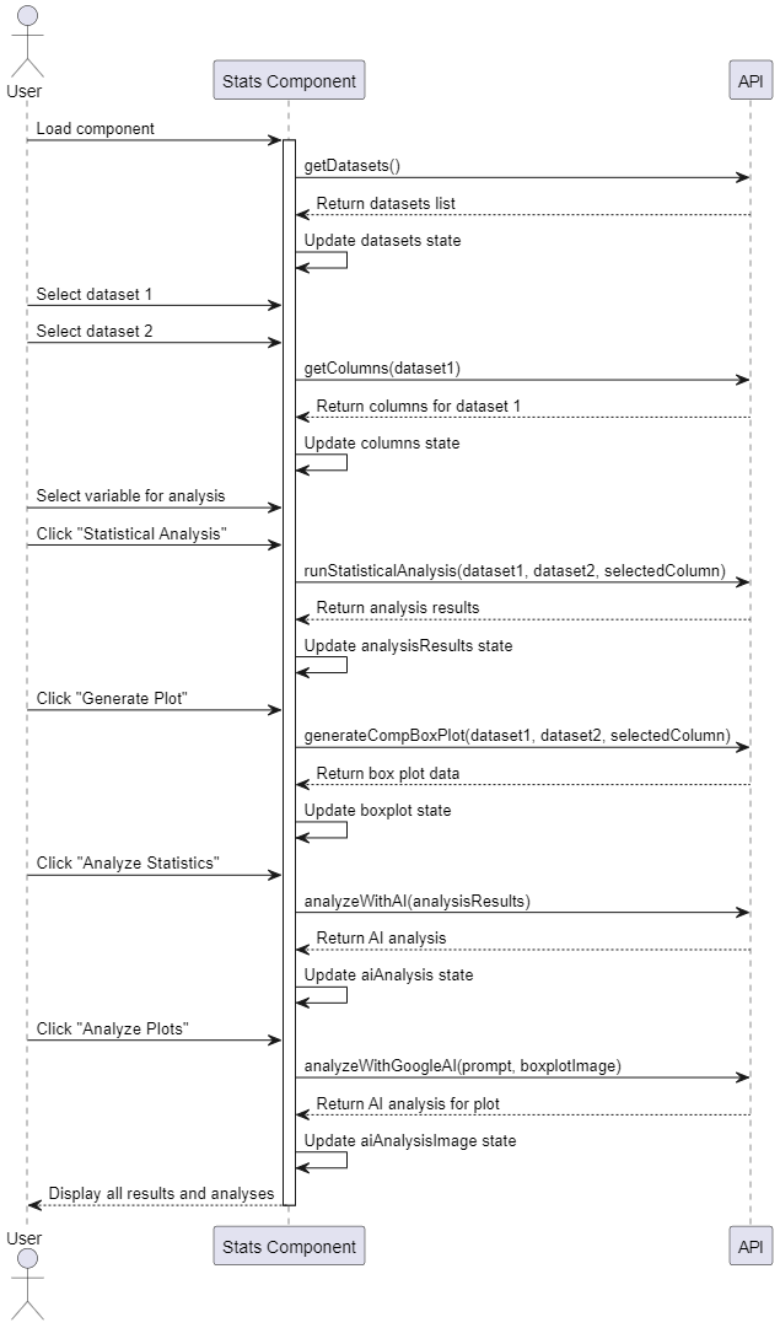


Nota. Diagrama de uso de la opción para generar y analizar estadísticas comparativas

En la Figura 20 vemos el diagrama de casos de uso de este componente y en la Figura 21 vemos el correspondiente diagrama de secuencia.

Figura 21

Diagrama de Secuencia - Generar Estadísticas Comparativas

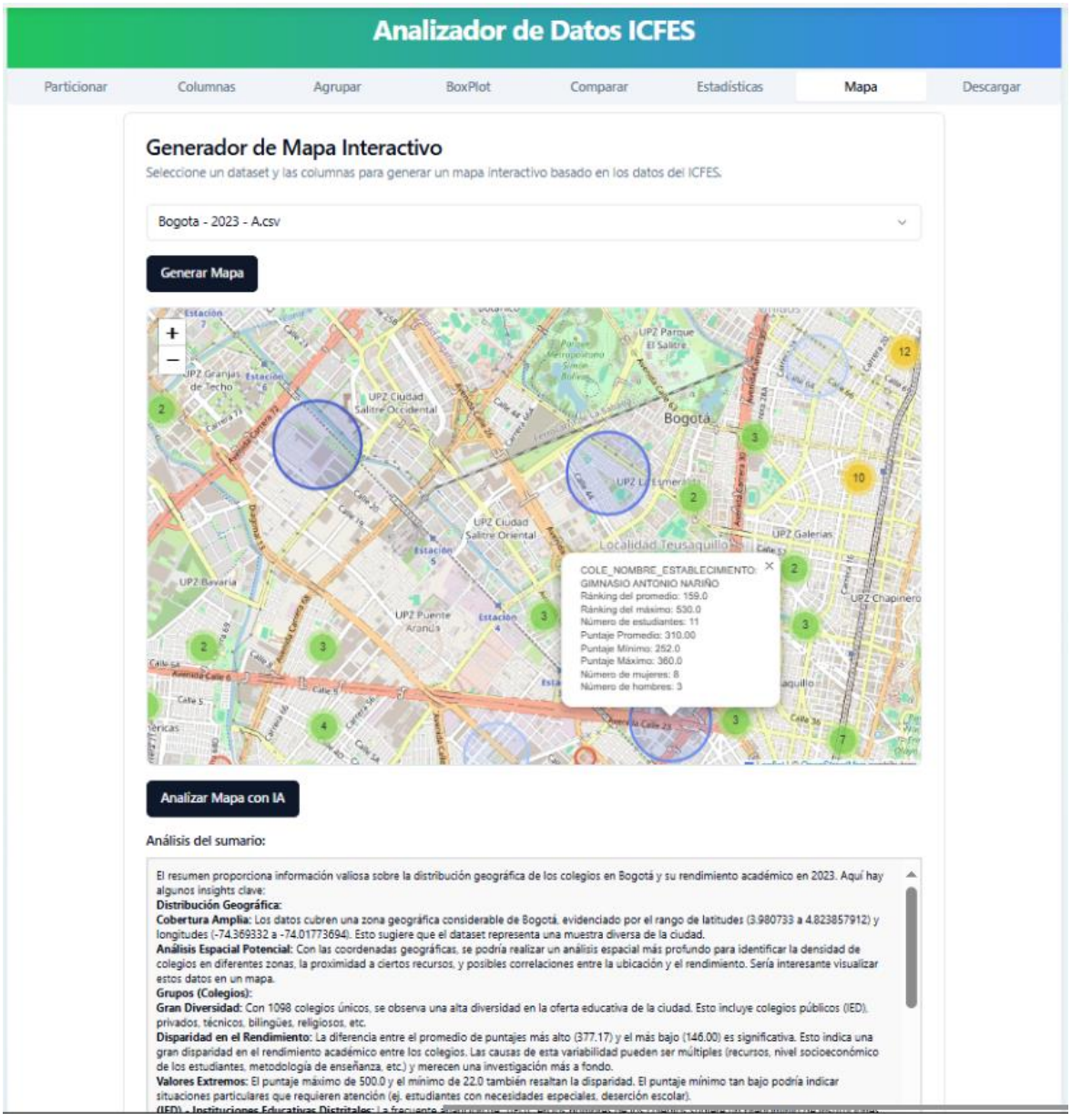


Nota. Secuencia de eventos para generar y analizar los comparativos entre diferentes sets de datos

Visualización de Mapas

Figura 22

Visualización en Mapa



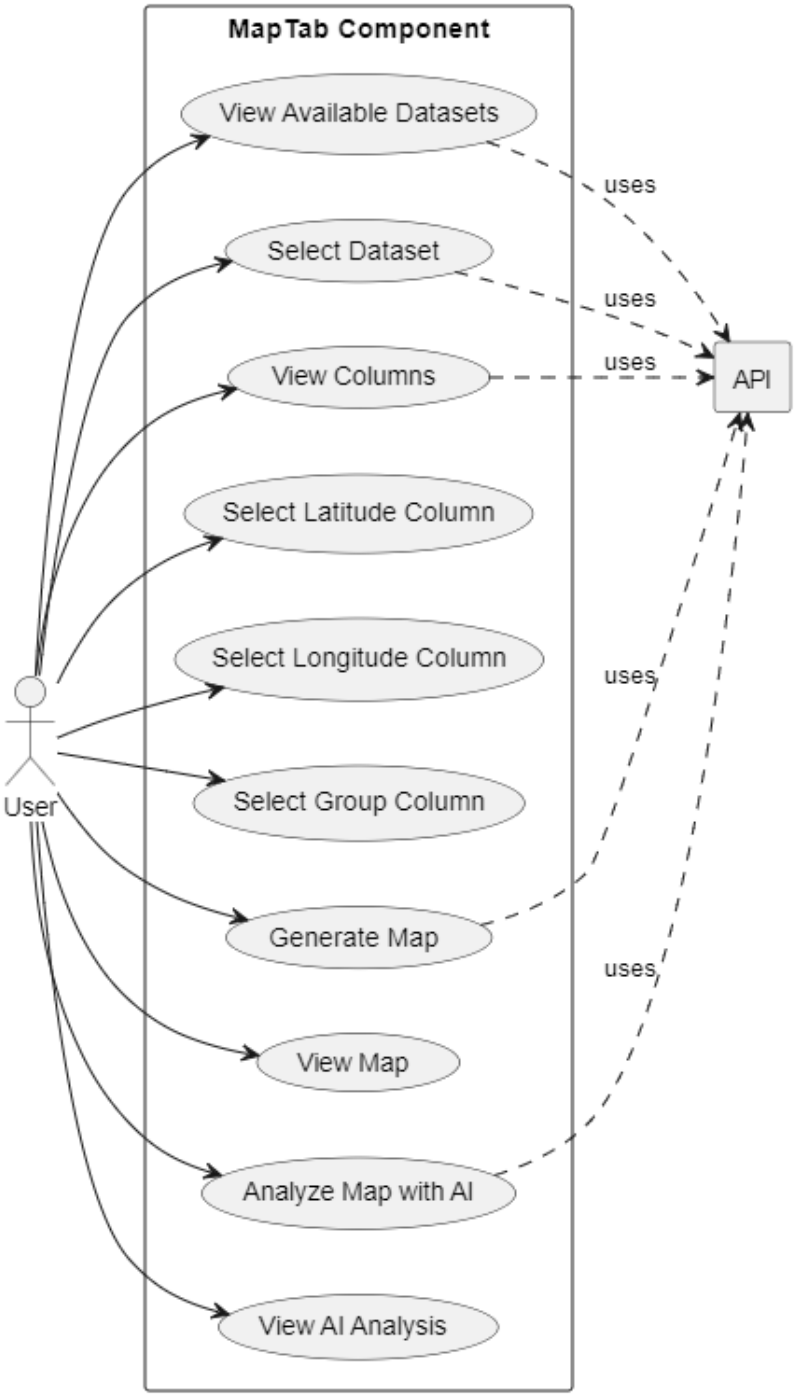
Nota. Pantalla de visualización de colegios del conjunto de datos en el mapa de Colombia

Se encuentra en la pestaña "Mapa". Permite a los usuarios visualizar datos geográficamente. EL usuario selecciona un conjunto de datos y la aplicación genera un mapa donde se presenta cada uno de los colegios en un círculo. Entre más grande y oscuro el círculo, será mejor el ranking del colegio en el conjunto de datos. Al dar click sobre alguno de los colegios, se presentan unos datos básicos como el puntaje promedio, máximo y mínimo de la institución, el número total de estudiantes que presentaron la prueba y el número de hombres y mujeres. En la Figura 22 se presenta el resultado del mapa para el conjunto de datos de Bogotá, calendario A.

En las siguientes figuras se aprecian los diagramas de casos de uso y de secuencia de la funcionalidad descrita anteriormente.

Figura 23

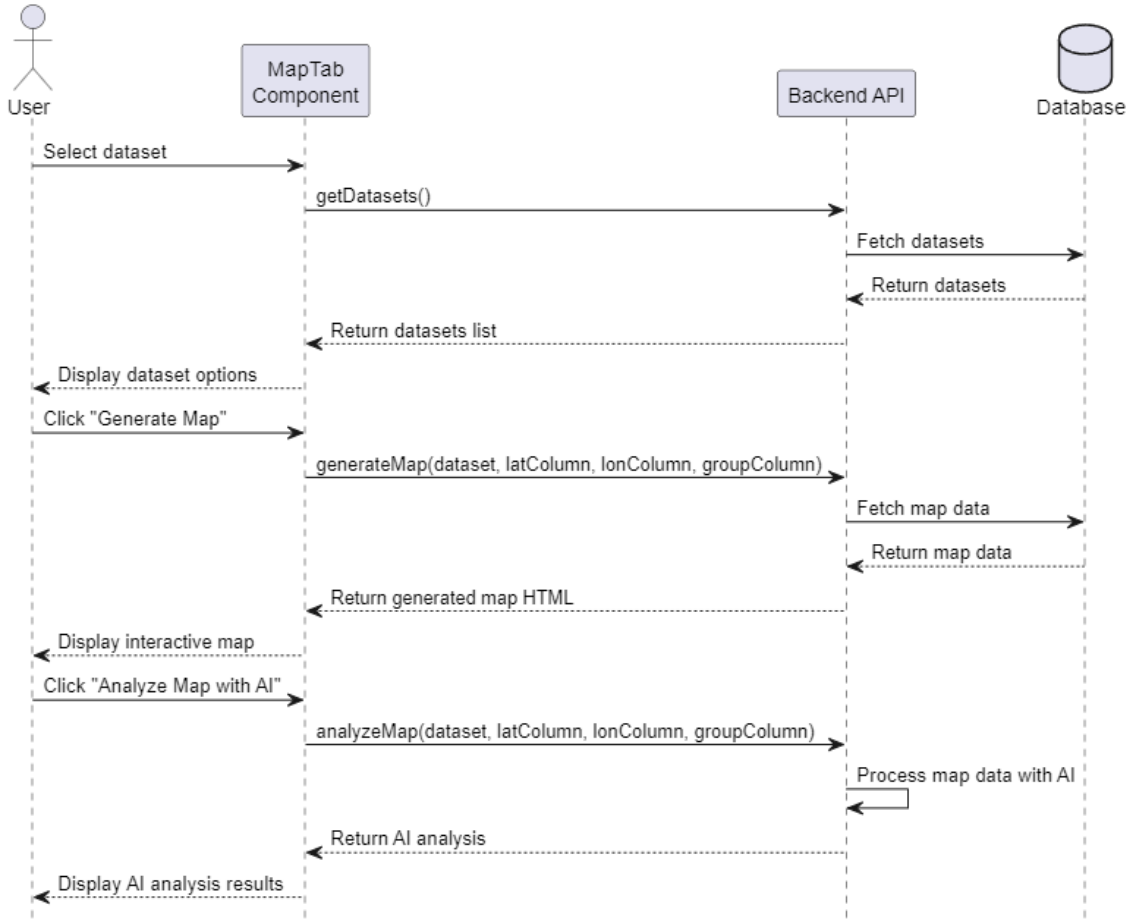
Caso de Uso - Generar Mapa



Nota. Diagrama de uso de la opción para visualizar los colegios del conjunto de datos en el mapa de Colombia

Figura 24

Diagrama de Secuencia - Generar Mapa



Nota. Diagrama de secuencia de la opción de generación de mapas

Descargar Datos

Accesible mediante la pestaña "Descargar". Permite a los usuarios exportar o descargar datos procesados para su uso fuera de la aplicación. El usuario selecciona el conjunto de datos, luego selección el formato del archivo de salida que puede ser CSV, JSON o Excel. Al oprimir el botón de salvar, se le pide al usuario la ubicación en su disco duro para descargar el archivo.

Figura 25

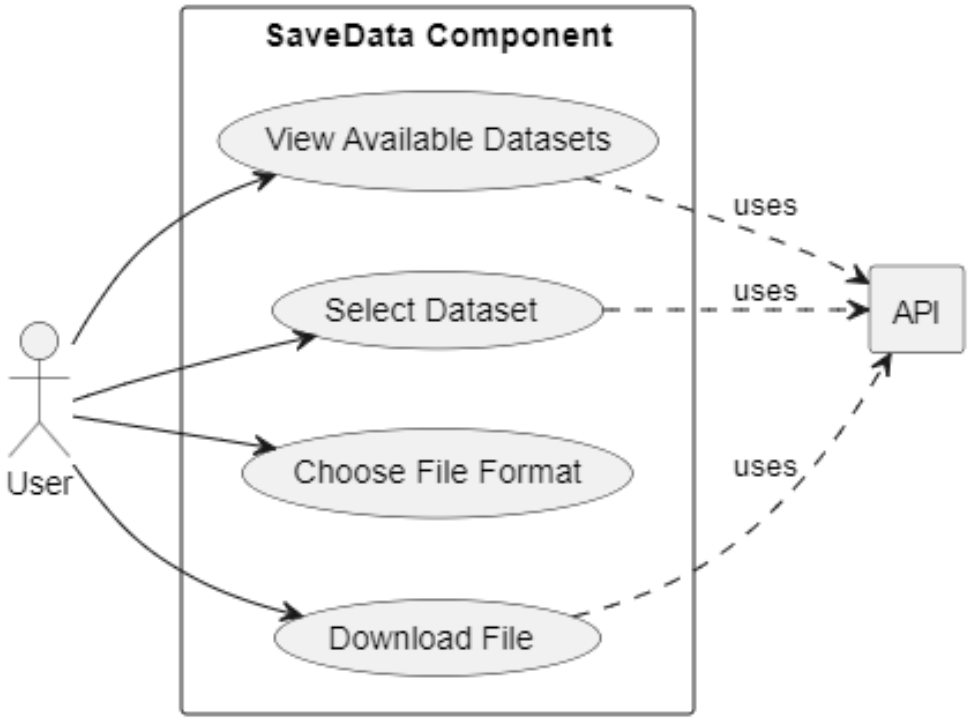
Descargar Datos

Nota. Pantalla del sistema para selección un conjunto de datos y descargarlo

A continuación, vemos los diagramas de casos de uso y de secuencia del componente

Figura 26

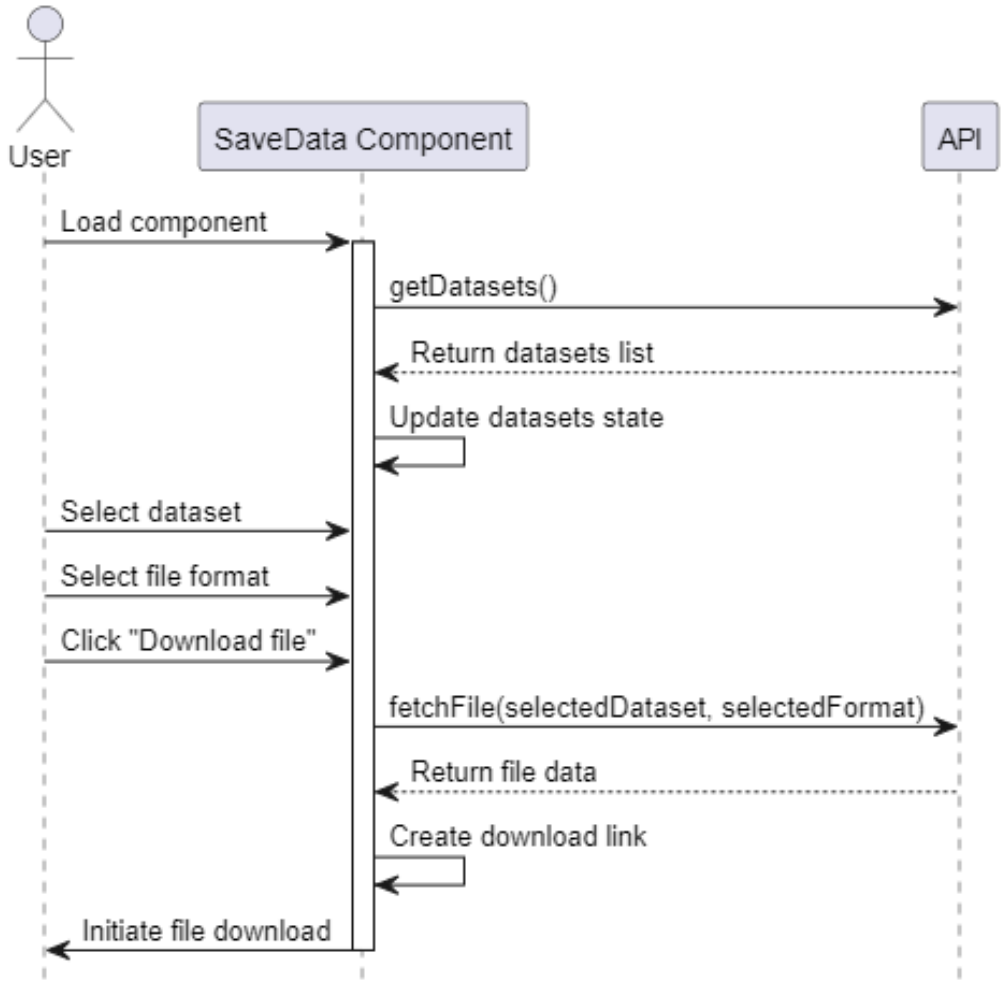
Caso de Uso - Descargar Datos



Nota. Diagrama de uso de la opción para descarga de conjuntos de datos al PC local.

Figura 27

Diagrama de Secuencia - Descargar Datos



Nota. Secuencia de eventos para descargar un conjunto de datos al PC local

Diagrama de Componentes

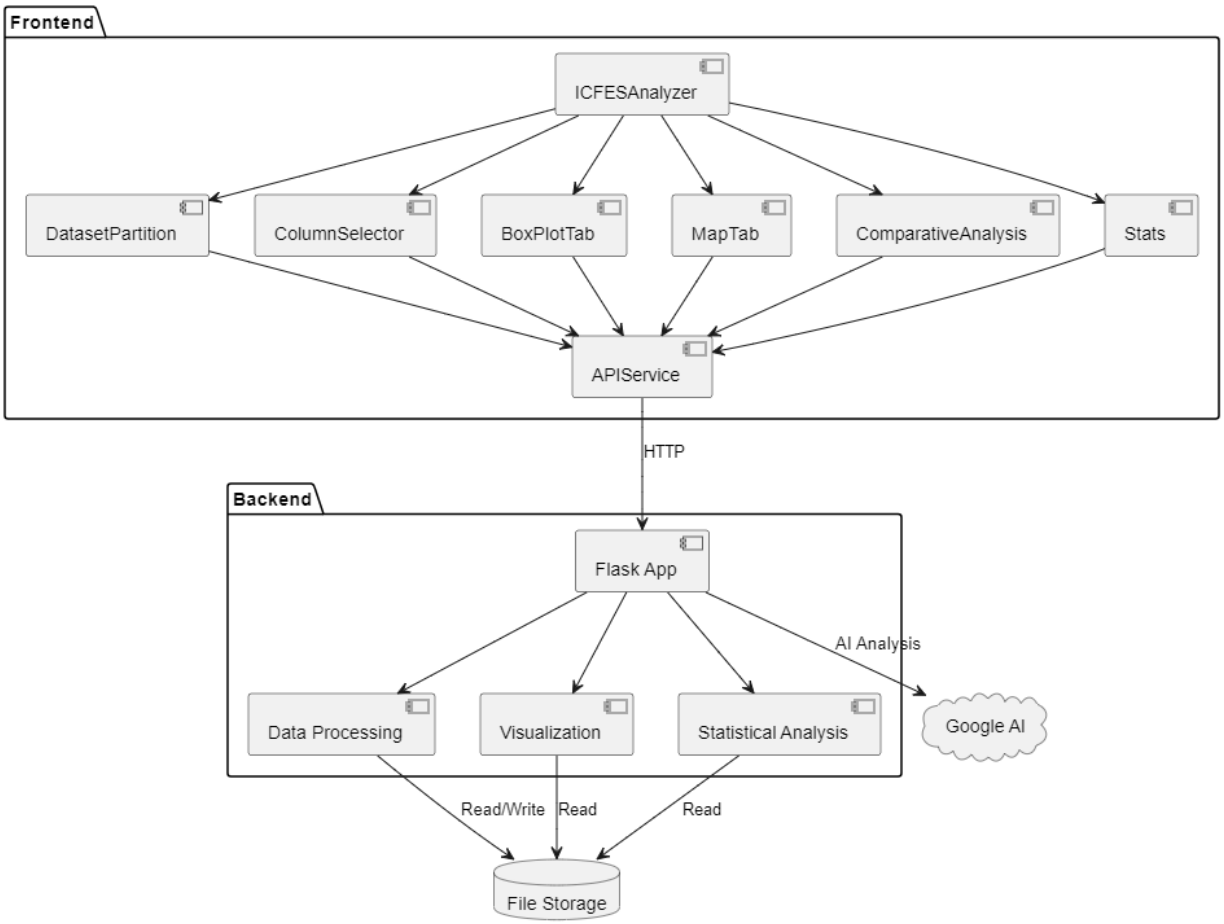
En la Figura 28 se aprecia el diagrama de componentes y los llamados a través de las diferentes capas de la aplicación.

Cada componente corresponde a una funcionalidad o caso de uso del sistema. Los componentes visuales del frontend se comunican con una API quien a su vez realiza los llamados

a la aplicación Flask del backend. A su vez, esta aplicación hace uso de librerías de funciones propias, de Python y de la API de Google AI para resolver las peticiones desde el f.rontend

Figura 28

Diagrama de Componentes



Nota. Diagrama que representa la interacción entre las capas y componentes del sistema

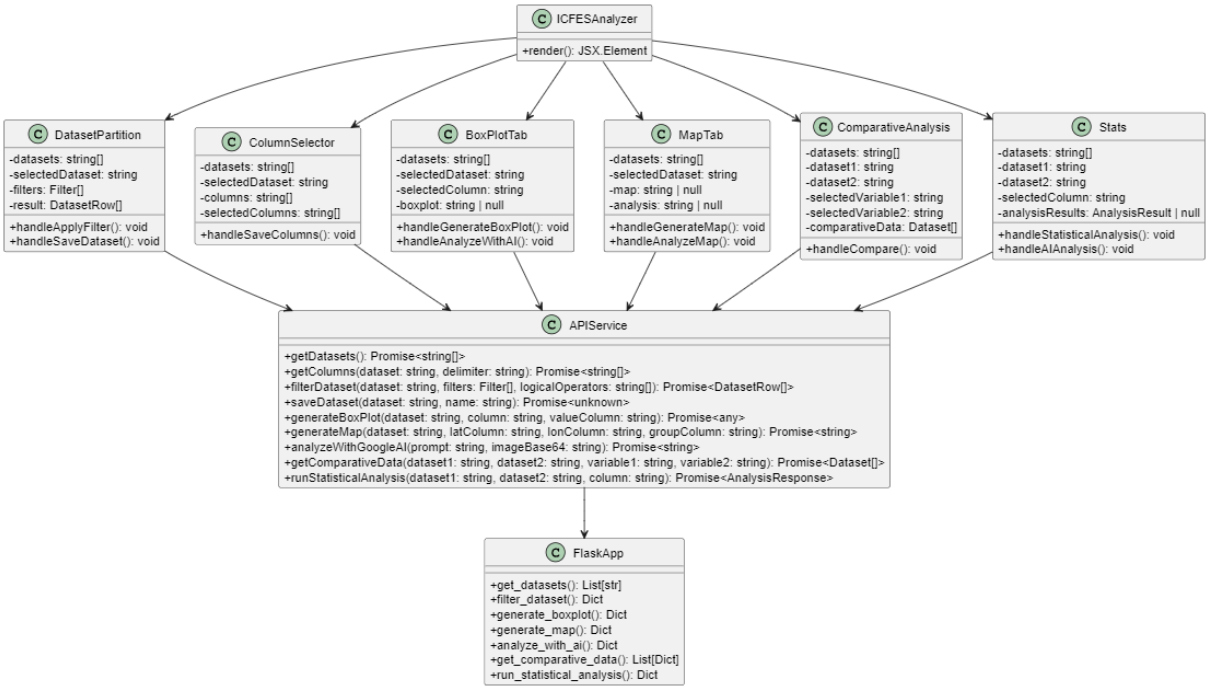
Diagrama de Clases

En la Figura 29 tenemos el diagrama de clases del sistema con cada uno de sus métodos y atributos. Cada clase es un ente que contiene funciones y atributos que se pueden acceder cuando

se generan instancias de estos. El diagrama presenta la interacción entre las diferentes clases del sistema.

Figura 29

Diagrama de Clases



Nota. Diagrama de los atributos y métodos de cada clase del sistema y la interacción entre ellos.

Conclusiones

El proyecto desarrollado proporciona a los investigadores y educadores en Colombia una herramienta efectiva para la revisión y selección inicial de los resultados de las pruebas ICFES Saber 11. Este sistema permite el acceso y análisis de los datos de manera sencilla y estructurada, superando las dificultades habituales asociadas a la dispersión de la información y su falta de organización. Durante el desarrollo de esta herramienta, se lograron implementar funcionalidades clave que han demostrado mejorar la capacidad de análisis y comprensión de los datos educativos a nivel nacional. A continuación, se destacan los logros principales:

Eficiencia en el Acceso y Análisis de Datos

Se centralizaron más de 5,541,529 observaciones correspondientes a estudiantes de todo el país entre 2014 y 2023. La centralización de los datos de ICFES en una única base de datos, junto con las capacidades de particionamiento y visualización, permite realizar análisis exhaustivos de los resultados de los estudiantes. Los datos ahora pueden dividirse de acuerdo con las necesidades del usuario, ya sea por años, colegios, municipios u otras variables de interés, facilitando estudios comparativos e históricos. La segmentación facilita estudios específicos y comparativos, permitiendo identificar patrones relevantes para diseñar políticas educativas diferenciadas.

Facilidad de Uso

Los investigadores sin conocimientos avanzados de programación se favorecen ya que el sistema fue diseñado con una interfaz amigable e intuitiva, eliminando la barrera técnica que existía para los investigadores no familiarizados con técnicas de programación. Esto ha permitido a los usuarios acceder a los datos y realizar análisis descriptivos y comparativos sin requerir habilidades técnicas avanzadas.

La IA hace el sistema accesible a una mayor diversidad de usuarios, democratizando el análisis de datos educativos. Investigadores, educadores y responsables de políticas, incluso sin experiencia en análisis de datos, pueden obtener interpretaciones confiables sin depender de expertos externos.

Visualización y Análisis Comparativo

La implementación de visualizaciones interactivas y la capacidad de generar estadísticas básicas y comparativas ofrecen una visión clara y completa de los datos. Las herramientas de visualización permiten a los investigadores identificar patrones y tendencias en los resultados de las pruebas Saber 11 del ICFES, facilitando la toma de decisiones y el diseño de políticas educativas informadas.

Pruebas de Normalidad

La incorporación de pruebas de normalidad, junto con un asistente de inteligencia artificial, aporta un valor añadido al sistema. Estas herramientas ofrecen una interpretación preliminar de los datos y sugieren posibles análisis adicionales, aumentando la eficiencia del proceso investigativo y mejorando la calidad del análisis de los datos obtenidos.

Potencial de Aplicación

El sistema fue diseñado con un enfoque flexible que permite su adaptación a otros conjuntos de datos relacionados con la educación en Colombia. Esta capacidad abre la posibilidad de un análisis más amplio de diferentes aspectos del sistema educativo colombiano, lo que puede contribuir a investigaciones en áreas relacionadas, como la infraestructura escolar, la tasa de deserción o el rendimiento en otros niveles educativos.

Interpretación Guiada y Contextualización de Resultados

La IA en el sistema ayuda a interpretar los resultados de los análisis de manera contextualizada, guiando a los usuarios sobre qué resultados son significativos o anómalos. Esta funcionalidad permite que investigadores con conocimientos limitados en estadística comprendan mejor los datos y sus implicaciones.

Generación Automática de Insights

El asistente de IA identifica patrones, relaciones y posibles correlaciones entre variables, generando insights que pueden pasar desapercibidos en un análisis manual. Esto enriquece el análisis, permitiendo obtener conclusiones más profundas y promoviendo una toma de decisiones mejor informada.

Recomendaciones

Es fundamental mantener la base de datos actualizada con los resultados de las pruebas ICFES de cada año. Para ello, se recomienda establecer una conexión directa con los portales de datos abiertos del gobierno, lo que permitiría automatizar el proceso de actualización de la información en el sistema.

Se recomienda explorar la aplicación de este proyecto en otros contextos educativos, como los resultados de otras pruebas estandarizadas o datos de censos escolares. Esto podría proporcionar una visión integral del estado del sistema educativo en Colombia y permitir análisis multivariables que incluyan factores de infraestructura, financiamiento, y variables demográficas.

Se recomienda considerar la integración con plataformas más avanzadas, como Power BI o Tableau, para permitir visualizaciones más complejas y personalizadas. Esto beneficiaría a usuarios con mayores requerimientos analíticos.

De igual manera sería beneficioso ofrecer una sección de capacitación o documentación adicional para guiar a los usuarios en el uso de todas las funcionalidades de la plataforma. Esto podría incluir tutoriales en video, guías interactivas y una sección de preguntas frecuentes.

Aunque el sistema está diseñado para manejar datos anonimizados, es importante considerar la inclusión de medidas adicionales de seguridad y privacidad, especialmente si se expandiera para abarcar datos que puedan contener información sensible. La implementación de capas de seguridad como autenticación y control de accesos, en consonancia con la normativa de protección de datos, podría ser una valiosa mejora.

Finalmente, se recomienda realizar estudios de impacto para evaluar cómo esta herramienta contribuye a la investigación educativa y al diseño de políticas en Colombia. Estos

estudios podrían enfocarse en medir el uso y la satisfacción de los usuarios, el alcance de las investigaciones realizadas con la herramienta y los cambios específicos en las políticas educativas derivados de su uso.

La asistencia de IA en el sistema no solo optimiza la accesibilidad y la comprensión de los resultados de los análisis, sino que también contribuye a que los usuarios generen conclusiones más profundas y significativas.

El proyecto desarrollado representa un avance significativo en la accesibilidad y análisis de datos educativos en Colombia. La herramienta cumple con su propósito de facilitar el análisis de los resultados de las pruebas ICFES Saber 11 y sienta una base sólida para futuras expansiones en el ámbito de la investigación educativa. Con las mejoras recomendadas, el sistema puede convertirse en una plataforma fundamental para investigadores, educadores y formuladores de políticas interesados en mejorar la calidad educativa en el país.

En la plataforma YouTube se encuentra un video con la demostración completa de la funcionalidad del prototipo (Leal, 2024).

Referencias Bibliográficas

AutoML | AutoML. (n.d.). Retrieved October 14, 2024, from https://www.automl.org/automl/

Bahamón, M., & Reyes, L. (2014). Characterization of the intellectual ability, Sociodemographic and academic factors of students with high and low performances in the Saber Pro exam – 2012. *Avances En Psicología Latinoamericana*, *32*(3), 459–476. https://doi.org/10.12804/apl32.03.2014.01

Congreso de Colombia. (2012). *Ley 1581 de 2012 - Gestor Normativo - Función Pública*. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Congreso de Colombia. (2014). *Ley 1712 de 2014 - Transparencia y del derecho de acceso a la información pública nacional*.

Data Studio now available as a Google Cloud service | Google Cloud Blog. (n.d.). Retrieved October 14, 2024, from https://cloud.google.com/blog/products/data-analytics/data-studio-now-available-as-a-google-cloud-service

Ferrer, J. M. (2010). *Evolución de los resultados de las pruebas Icfes y saber en las áreas de matemática y lenguaje en Risaralda - Evolution of the Icfes and Saber test results in the Mathematics and Language areas in Risaralda* (Issue 88). https://dialnet.unirioja.es/servlet/articulo?codigo=4897812

Getting started with RapidMiner Studio - Altair RapidMiner Documentation. (n.d.). Retrieved October 14, 2024, from https://docs.rapidminer.com/9.9/studio/getting-started/index.html

H2O.ai. (2024). *H2O.ai | Convergence of The World's Best Predictive & Generative AI*. https://h2o.ai/

ICFES. (2024a). *Acerca del examen Saber 11° – ICFES*. http://129.159.121.61/evaluaciones-icfes/acerca-del-examen-saber-11/

ICFES. (2024b). *Oficina de investigaciones - 01. DataIcfes - Todos los documentos*. https://icfesgovco.sharepoint.com/sites/Oficinadeinvestigaciones/Documentos%20compartidos/Forms/AllItems.aspx?id=%2Fsites%2FOficinadeinvestigaciones%2FDocumentos%20compartidos%2F01%2E%20DataIcfes&p=true&ga=1

Leal, J. (2024). *Demostración aplicativo de visualización de resultados ICFES - YouTube*. https://www.youtube.com/watch?v=WAyHjMUJD5c&ab_channel=JohnLeal

López Murphy Y, J. J., & Zarza, G. (2017). La Ingeniería del Big Data: Como Trabajar Con Datos. In *Print. Tecnología* (Vol. 534). https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052020000400556

Matplotlib — Visualization with Python. (n.d.). Retrieved October 14, 2024, from https://matplotlib.org/

Ministerio de Tecnologías de la Información y las Comunicaciones. (2021). *Guía para el uso y aprovechamiento de Datos Abiertos en Colombia*. https://herramientas.datos.gov.co/sites/default/files/2021-08/Guia%20de%20Datos%20Abiertos%20de%20Colombia.pdf

MINTIC. (2024). *Datos Abiertos Colombia | Datos Abiertos Colombia*. https://www.datos.gov.co/

Open for Innovation | KNIME. (n.d.). Retrieved October 14, 2024, from https://www.knime.com/

Plotly Python Graphing Library. (n.d.). Retrieved October 14, 2024, from https://plotly.com/python/

Power BI - Data Visualization | Microsoft Power Platform. (n.d.). Retrieved October 14, 2024, from https://www.microsoft.com/en-us/power-platform/products/power-bi

Presidencia de la República. (2015). *Decreto 103 de 2015 Nivel Nacional*. https://www.bogotajuridica.gov.co/sisjur/normas/Norma1.jsp?i=60556

Project Jupyter | Home. (n.d.). Retrieved October 14, 2024, from https://jupyter.org/

¿Qué es Tableau? (n.d.). Retrieved October 14, 2024, from https://www.tableau.com/es-es/why-tableau/what-is-tableau

Rosero, D. D. R., Ortega, R. E. O., & Villota, M. E. H. (2021). Academic performance determinants of high school students in the Department of Nariño, Colombia. *Lecturas de Economia*, *94*, 87–126. https://doi.org/10.17533/UDEA.LE.N94A341834

Shang, Z., Zgraggen, E., Buratti, B., Kossmann, F., Eichmann, P., Chung, Y., Binnig, C., Upfal, E., & Kraska, T. (2019). Democratizing data science through interactive curation of ML pipelines. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1171–1188. https://doi.org/10.1145/3299869.3319863

Software IBM SPSS. (n.d.). Retrieved October 14, 2024, from https://www.ibm.com/es-es/spss

Statistical software for data science | Stata. (n.d.). Retrieved October 14, 2024, from https://www.stata.com/

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021