



UNIVERSIDAD SANTO TOMÁS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
T U N J A

TÍTULO PASANTÍA

Dashboard interactivo para la analítica y visualización de resultados Saber Pro en la USTA

PROPONENTE(S)

Christian Alberto Montes Patiño

1049657211

2220428

DIRECTOR

David Gerónimo Soler Gómez

Tunja

01/06/2026

CONTENIDO

CONTENIDO.....	2
ÍNDICE DE TABLAS.....	4
ÍNDICE DE FIGURAS.....	4
FICHA TÉCNICA DE LA PASANTÍA.....	5
ANTECEDENTES DE LA EMPRESA.....	7
PLANTEAMIENTO DEL PROBLEMA.....	8
JUSTIFICACIÓN.....	10
OBJETIVOS.....	11
Objetivo General.....	11
Objetivos específicos.....	11
MARCO TEÓRICO.....	12
Inteligencia de Negocios (Business Intelligence - BI) y Dashboards.....	12
Proceso ETL (Extracción, Transformación y Carga).....	12
Modelado Predictivo y Estudio de Viabilidad.....	12
Expresiones Regulares (Regex / RE).....	13
DAX (Data Analysis Expressions).....	13
Microdatos.....	13
Núcleo Básico de Conocimiento (NBC).....	13
Percentil.....	13
Examen de Estado.....	14
Pruebas Saber 11.....	14
Pruebas Saber Pro.....	14
Diccionario de Datos.....	15
MODELO DE DESARROLLO.....	16
Entendimiento Institucional y de Contexto (Comprensión del negocio).....	17
Diagnóstico y Exploración de Fuentes (Comprensión de los datos).....	17
Ingeniería y Adecuación de Datos (Preparación de los datos).....	17
Análisis de Patrones e Indicadores Críticos (Modelado conceptual y analítico).....	18
Evaluación de Viabilidad Predictiva y Estándares (Evaluación de viabilidad y calidad)...	18
Estructuración y Construcción del Dashboard Web (Implementación).....	18
Aseguramiento, Pruebas y Despliegue (Retroalimentación continua).....	18
DESARROLLO DEL PROYECTO.....	25
Fase 1. Gestión de datos.....	31

Recolección de datos.....	31
Limpieza e integración.....	35
Validación de datos integrados.....	41
Fase 2. Análisis y viabilidad.....	45
Análisis descriptivo y exploratorio.....	45
Visualización preliminar.....	48
Documentación de hallazgos iniciales.....	51
Revisión de estándares ICFES.....	53
Estudio de viabilidad técnica.....	55
Documentación de conclusiones y recomendaciones.....	58
Fase 3. Desarrollo y despliegue.....	59
Diseño de Mockup y arquitectura del Dashboard.....	59
Desarrollo del Dashboard Web.....	64
Pruebas funcionales y despliegue.....	65
CONCLUSIONES.....	68
Cumplimiento de objetivos.....	68
(Objetivo específico 1).....	68
(Objetivo específico 2).....	69
(Objetivo específico 3).....	69
(Objetivo específico 4).....	70
Retos y dificultades superadas.....	71
Aprendizaje y competencias desarrolladas.....	72
Impacto, automatización y beneficios concretos.....	73
Proyección institucional.....	73
RECOMENDACIONES Y TRABAJOS FUTUROS.....	74
ANEXOS.....	75
REFERENCIAS.....	76

ÍNDICE DE TABLAS

Tabla 1. Ficha técnica.....	5
Tabla 2. Actividades desarrolladas y condiciones técnicas.....	18
Tabla 3. Fechas de actividades desarrolladas y productos entregados.....	24
Tabla 4. Software, los lenguajes y librerías utilizadas.....	28
Tabla 5. Retos técnicos y metodológicos.....	70

ÍNDICE DE FIGURAS

Figura 1 Diagrama de pescado. Causas y consecuencias.....	9
Figura 2 Diagrama de procesos, modelo de desarrollo.....	16
Figura 3 Diagrama de actividades y secuencia de flujo.....	24
Figura 4 Imagen tomada del Formulario Datalcfes.....	32
Figura 5 Imagen tomada del Documento Datalcfes.....	32
Figura 6 Imagen tomada en Sharepoint, directorio Datalcfes.....	33
Figura 7 Imagen tomada de los archivos descargados de Datalcfes (.zip).....	34
Figura 8 verificación de separadores en archivos .txt.....	36
Figura 9 Normalización de formatos.....	36
Figura 10 Estandarización de valores vacíos/nulos.....	37
Figura 11 Estudiantes que presentaron su examen saber Pro en el año 2018.....	38
Figura 12. Imagen recortada del uso de expresiones regulares en la extracción de los datos.....	40
Figura 13. Tiempo promedio de análisis por PDFs (1212): 1.30s por archivo.....	40
Figura 14 Cruce Saber Pro con Saber 11.....	41
Figura 15 Resultados del cruce, distribuido por NBC.....	42
Figura 16. Explicación niveles de desempeño para el módulo de lectura crítica (Saber Pro) 46	
Figura 17 Sección NBC a través del tiempo (Filtro NBC y Sedes Usta).....	48
Figura 18. Sección Percentil por NBC (Filtro Sedes Usta).....	49
Figura 19 Sección desempeño por competencia (Filtro Nivel de agregación y Año).....	49
Figura 20 Sección desempeño por NBC (Filtro Sedes Usta y Año).....	50
Figura 21. Matriz de evaluación estratificada por Núcleo Básico del Conocimiento (NBC)56	
Figura 22. Borrador en Figma.....	59
Figura 23 Sección NBC a través del tiempo (Filtro NBC y Sedes Usta), en producción....	60
Figura 24 Sección Percentil por NBC (Filtro Sedes Usta), en producción.....	61

Figura 25 Sección desempeño por competencia (Filtro Nivel de agregación y Año), en producción.....	62
Figura 26 Sección desempeño por NBC (Filtro Sedes Usta y Año), en producción.....	63
Figura 27. Código Dax para la creación de los niveles de agregación.....	64
Figura 28. Código Dax para la creación del filtro desempeño.....	64
Figura 29. Publicar Dashboard desde el aplicativo local Desktop al aplicativo Web.....	65
Figura 30. Buscar Dashboard en el aplicativo web.....	66
Figura 31. Publicar Dashboard desde el aplicativo Web.....	66
Figura 32. Muestra del enlace generado cuando el Dashboard se publica en la Web.....	67

FICHA TÉCNICA DE LA PASANTÍA

Tabla 1. Ficha técnica

Título	Dashboard interactivo para la analítica y visualización de resultados Saber Pro en la USTA.
Nombre Estudiante	Christian Alberto Montes Patiño.
Documento estudiante	1049657211
Correo electrónico estudiante	Christian.montes@usantoto.edu.co
Director	David Gerónimo Soler Gómez
Entidad o sector donde realizará la pasantía	Universidad Santo Tomás, Seccional Tunja
Lugar de ejecución de la pasantía	Universidad Santo Tomás, Seccional Tunja
Duración	Cuatro meses.

Los abajo firmantes confirman que todos los datos incluidos en la presente propuesta son correctos y verídicos, que no incumplen ninguna ley o norma vigente (incluir nombres y firmas de estudiantes y director).

Firma del autor
Nombre autor

Firma del director
Nombre director pasantía

ANTECEDENTES DE LA EMPRESA

La Universidad Santo Tomás, como la primera Universidad de Colombia, ha encabezado un compromiso histórico y firme con el mejoramiento continuo de la calidad educativa y el aseguramiento de estándares académicos de excelencia. La institución participa activamente en los procesos de evaluación establecidos por el Estado, incluyendo las Pruebas Saber Pro, como parte de su metodología. La Dirección de Aseguramiento de la Calidad de la Universidad Santo Tomás – Seccional Tunja forma parte de esta metodología, fomentando la excelencia en los procesos de evaluación institucional, acreditación y mejoramiento continuo de los programas académicos. Históricamente, este departamento ha gestionado los resultados de las Pruebas Saber Pro mediante la recopilación manual de información desde la plataforma de datos abiertos del ICFES y los reportes individuales de los estudiantes. Los análisis posteriores se han realizado en hojas de cálculo, mediante métodos estadísticos, generando respectivos reportes semestrales para las diferentes áreas académicas.

Con el incremento constante en el número de estudiantes y la necesidad de aprovechar el valor de la información que conllevan los datos, se ha previsto que los métodos tradicionales de procesamiento y análisis resultan insuficientes para las demandas actuales por parte de directivos, coordinadores de programa y docentes, y de la visión estratégica que ha definido la Dirección de Aseguramiento de la Calidad.

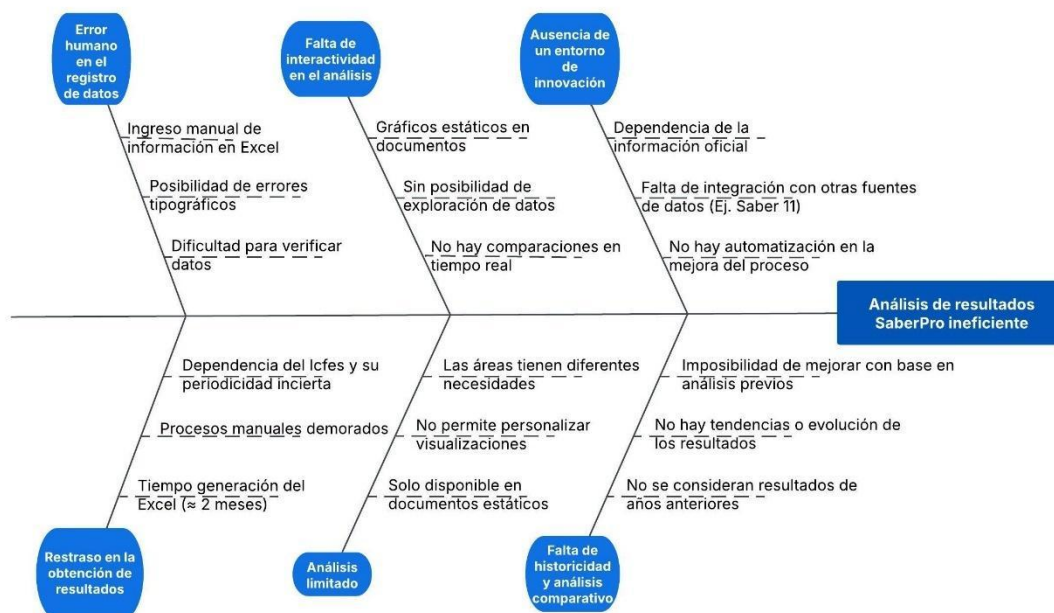
PLANTEAMIENTO DEL PROBLEMA

Cada año a nivel nacional, los estudiantes universitarios que han aprobado al menos el 75% de sus créditos académicos presentan un examen de Estado denominado “Pruebas Saber Pro”, que pone a prueba competencias a lo largo de su estancia académica, tanto generales como especializadas. La universidad Santo Tomás, y las demás universidades en Colombia, tienen acceso a los resultados desde la plataforma de datos abiertos (Icfes.gov.co), o proporcionada por los propios estudiantes mediante archivos (.PDF) que muestran sus informes individuales. Desafortunadamente, los resultados mantienen un rezago de un año desde la plataforma de datos abiertos, y 3-4 meses por parte de los estudiantes; sin contar el tiempo de trabajo manual invertido para procesar la información obtenida. Por lo anterior, existe una necesidad creciente por parte de los directivos y docentes de acceder a dicha información y su analítica respectiva con suma prontitud.

La ausencia de una herramienta de visualización y análisis de datos obliga a depender de métodos manuales, como la organización y filtrado de datos en Excel, seguido de la generación de gráficos. Este proceso toma varias semanas o meses, es propenso a errores y restringe la maleabilidad de los datos, limitando la manera de extraer información de valor. La falta de una plataforma que pueda automatizar y centralizar los datos para el análisis, impide la identificación oportuna de áreas de mejora en el rendimiento de los estudiantes de la Universidad Santo Tomás, desaprovechando así una excelente base empírica para una toma de decisiones más informadas y la implementación de estrategias efectivas para el mejoramiento continuo de la calidad educativa.

El siguiente diagrama de espina de pescado [Figura 1] nos muestra un resumen conciso de las causas de la situación problema.

Figura 1 Diagrama de pescado. Causas y consecuencias.



Fuente: Autor

JUSTIFICACIÓN

La pertinencia de este proyecto se fundamenta en la necesidad imperativa que tiene la Dirección de Aseguramiento de la Calidad de la Universidad Santo Tomás de optimizar la toma de decisiones estratégicas mediante la transición de un manejo manual, descentralizado y propenso a errores de los microdatos de las pruebas Saber Pro, hacia un sistema de analítica institucional centralizado y automatizado. La implementación de esta solución se justifica al mitigar los reprocesos operativos actuales y, simultáneamente, proyectar un impacto significativo a mediano y largo plazo en la gestión universitaria; por un lado, proveerá a las decanaturas y comités curriculares de una herramienta ágil para identificar brechas en competencias genéricas y diseñar planes de mejoramiento pedagógico oportunos antes de las auditorías oficiales, y por el otro, establecerá una infraestructura de datos normalizada y escalable que servirá como base metodológica para futuros estudios de analítica predictiva y medición del valor agregado mediante el cruce relacional con bases de datos históricas de Saber 11. La ejecución de este proyecto es fundamental para modernizar los mecanismos de auditoría académica en la Universidad Santo Tomás, transitando de un esquema de datos estáticos hacia una cultura de analítica institucional. El cumplimiento de los objetivos principales permitirá entregar soluciones concretas, herramientas que democratizan el acceso a la información, permitiendo a las decanaturas e investigadores explorar indicadores multidimensionales en tiempo real para diseñar estrategias pedagógicas efectivas y respaldar los procesos de acreditación de alta calidad.

OBJETIVOS

Objetivo General

Desarrollar una solución tecnológica para el análisis y visualización de los resultados Saber Pro de la Universidad Santo Tomás, integrando un dashboard interactivo y un estudio de viabilidad de modelos predictivos, para fortalecer la toma de decisiones académicas desde la Dirección de Aseguramiento de la Calidad.

Objetivos específicos

1. Integrar los datos públicos de las Pruebas Saber Pro y los datos institucionales mediante procesos de limpieza, validación y estructuración, para obtener un conjunto de datos unificado y documentado listo para análisis.
2. Analizar los resultados históricos de las Pruebas Saber Pro mediante técnicas de análisis estadístico descriptivo y exploratorio, para identificar patrones y brechas de desempeño académico por programa y cohorte.
3. Realizar un estudio de viabilidad técnica para la implementación de un modelo predictivo sobre los resultados Saber Pro, mediante el análisis de los estándares ICFES y los datos institucionales disponibles, para determinar las condiciones y alcances de su aplicación por programa académico.
4. Implementar un Dashboard interactivo de visualización y análisis de datos mediante Power BI, para centralizar el acceso a los resultados Saber Pro y apoyar la toma de decisiones de la Dirección de Aseguramiento de la Calidad.

MARCO TEÓRICO

Inteligencia de Negocios (Business Intelligence - BI) y Dashboards

La Inteligencia de Negocios agrupa los procesos tecnológicos que recopilan, organizan y analizan los datos de una institución para generar información de valor. Un *dashboard* o tablero de control interactivo es una herramienta visual derivada del BI, diseñada para centralizar indicadores clave en gráficas comprensibles. Esto permite a los directivos visualizar reportes históricos en tiempo real sin requerir conocimientos técnicos.

Proceso ETL (Extracción, Transformación y Carga)

ETL es una metodología de arquitectura de datos utilizada para automatizar el manejo de grandes volúmenes de información. El proceso consta de tres fases: *Extracción* (obtener datos desde su origen, como archivos PDF estructurados), *Transformación* (limpiar, estandarizar formatos y corregir errores) y *Carga* (almacenar la información lista en un sistema centralizado).

Modelado Predictivo y Estudio de Viabilidad

El modelado predictivo es una técnica de análisis de datos que evalúa información histórica para proyectar escenarios, comportamientos o resultados futuros. En proyectos de infraestructura de datos, antes de implementar modelos predictivos, es un requisito estricto ejecutar un estudio de viabilidad técnica. Esto permite auditar si el volumen, el equilibrio estadístico y la calidad del histórico son aptos. Al determinarse que la implementación algorítmica y el modelado predictivo no eran viables en el escenario inicial, la estrategia prioriza la consolidación de los datos mediante el desarrollo de un *dashboard* analítico funcional, dejando evaluada la viabilidad estadística futura.

Expresiones Regulares (Regex / RE)

Las expresiones regulares son patrones de búsqueda programados para identificar y extraer combinaciones precisas de texto dentro de un documento. En el contexto de la minería de texto, permiten escanear cientos de reportes no estructurados para capturar selectivamente únicamente los datos de interés (como números de cédula o puntajes específicos) de forma automatizada y rápida.

DAX (Data Analysis Expressions)

DAX es el lenguaje de fórmulas y expresiones utilizado por motores analíticos de Microsoft, como Power BI. A diferencia de fórmulas estáticas (como las de Excel), DAX permite calcular y recalcular métricas de manera dinámica basándose en los filtros y contextos de navegación que el usuario final seleccione en la pantalla.

Microdatos

Los microdatos representan el nivel más básico y granular de la información. Hacen referencia a las bases de datos crudas, estructuradas a nivel de individuo o de registro único (por ejemplo, el resultado exacto de un solo estudiante), antes de ser agrupados en promedios o estadísticas generales.

Núcleo Básico de Conocimiento (NBC)

Es una clasificación oficial implementada por el Estado colombiano para agrupar los diferentes programas académicos de educación superior basándose en sus disciplinas, áreas o campos profesionales afines (por ejemplo, agrupar las ciencias económicas o las diferentes vertientes de ingeniería).

Percentil

Es una medida estadística de posición no central que divide un grupo de datos ordenados en cien partes porcentualmente iguales. En el entorno académico, indica la posición

relativa de un estudiante frente a un grupo. Por ejemplo, si un resultado se ubica en el percentil 85, significa que dicho puntaje supera al 85% de la población evaluada.

Examen de Estado

Un Examen de Estado es un instrumento estandarizado de evaluación externa diseñado y aplicado en Colombia por el Instituto Colombiano para la Evaluación de la Educación (ICFES). Su propósito principal es medir y monitorear la calidad de la educación formal impartida en el país en sus diferentes niveles (Ministerio de Educación Nacional, s.f.). Los resultados de estas pruebas sirven como indicador estadístico fundamental para el aseguramiento de la calidad institucional y el diseño de políticas públicas.

Pruebas Saber 11

Es el Examen de Estado orientado a evaluar el nivel de desarrollo de las competencias de los estudiantes que están por culminar la educación media (grado undécimo). Evalúa de manera transversal áreas como Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés (ICFES, s.f.-a). El ICFES suele aplicar esta prueba dos veces al año (adaptándose a los calendarios académicos A y B de los colegios) y constituye el requisito estándar para el ingreso a la educación superior en Colombia (ICFES, s.f.-a).

Pruebas Saber Pro

Anteriormente conocidas como ECAES, las Pruebas Saber Pro son el Examen de Estado obligatorio para los estudiantes de educación superior que han aprobado al menos el 75% de los créditos académicos de sus programas de pregrado universitario (ICFES, s.f.-b). Evalúan tanto competencias genéricas (razonamiento cuantitativo, lectura crítica, competencias ciudadanas, comunicación escrita e inglés) como competencias específicas según el Núcleo Básico de Conocimiento (NBC) de la carrera (ICFES, s.f.-b).

Generalmente, el ICFES dispone de dos aplicaciones anuales para facilitar la cobertura de las diferentes cohortes próximas a graduarse.

Diccionario de Datos

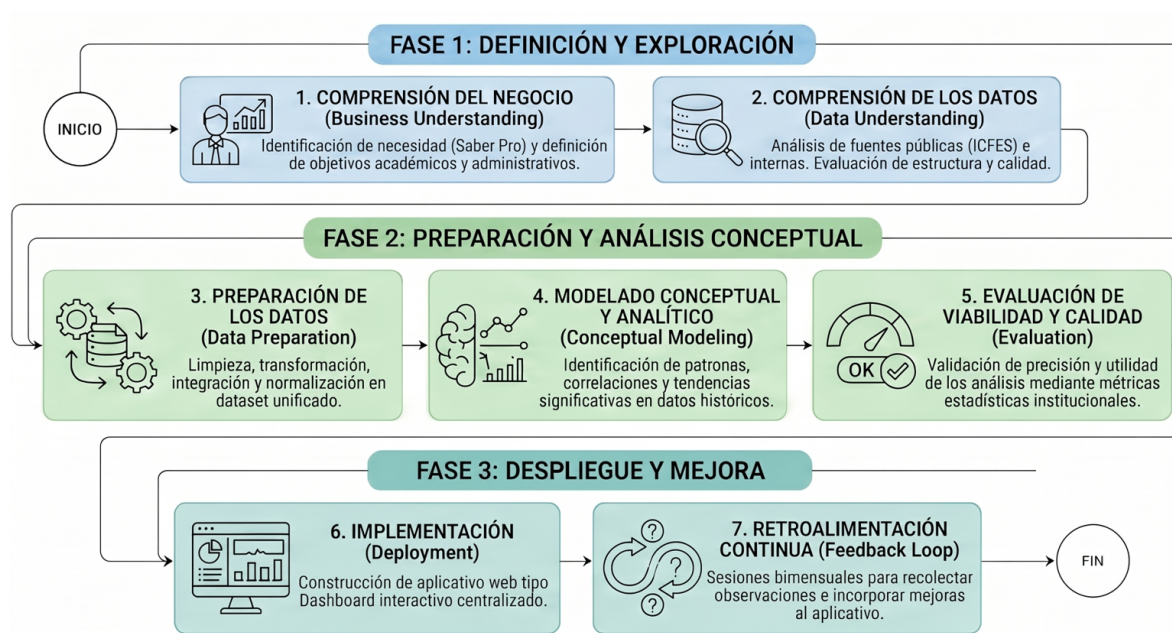
Un diccionario de datos es un documento o repositorio centralizado que detalla la estructura, el significado, el formato y las características de los datos utilizados en un sistema o proyecto (IBM, s.f.). Funciona como un catálogo o "mapa" que permite a cualquier usuario comprender qué representa exactamente cada variable o columna dentro de una base de datos compleja. En el contexto de la estructuración de información institucional, es una herramienta fundamental para estandarizar la semántica de los datos crudos, asegurando que la interpretación de métricas y categorías no dependa de suposiciones, sino de reglas claras y documentadas (Gartner, s.f.).

MODELO DE DESARROLLO

Las fases que conformaron el modelo de desarrollo de la pasantía se caracterizaron por un lenguaje claro, una orientación práctica institucional y una secuencia lógica de hitos alcanzados.

En la Figura 2 se muestra cómo el modelo de desarrollo del proyecto se dividió en las diferentes fases, alineadas con los componentes del modelo metodológico.

Figura 2 Diagrama de procesos, modelo de desarrollo.



Fuente: Autor

Como podemos ver en la Figura 2, la metodología aplicada en el modelo de desarrollo consistió en siete (7) procesos, distribuidos en tres (3) fases, en las cuáles se evidenció la finiquitación de los objetivos propuestos. Cada proceso tuvo un objetivo y alcance claro, tal y como se va a explicar abajo.

Entendimiento Institucional y de Contexto (Comprensión del negocio)

Esta fase inicial consistió en delimitar las necesidades de las áreas académicas y administrativas de la Universidad. Se identificó que la prioridad institucional radicaba en disponer de un mecanismo ágil para centralizar, auditar y explorar de manera interactiva el histórico de los resultados Saber Pro, facilitando así la toma de decisiones curriculares y la detección temprana de oportunidades de mejora.

Diagnóstico y Exploración de Fuentes (Comprensión de los datos)

Se procedió al reconocimiento y recolección de los conjuntos de datos. Esto incluyó la descarga de microdatos públicos provistos por el Estado (ICFES) correspondientes al periodo temporal disponible (Saber Pro y Saber 11), así como la consolidación de los registros académicos individuales más recientes (PDF de resultados individuales de las pruebas Saber Pro) gestionados internamente por la institución educativa.

Ingeniería y Adecuación de Datos (Preparación de los datos)

Fase crítica enfocada en la preparación del insumo analítico. Se unificaron las bases de datos dispersas mediante llaves lógicas de conexión y se ejecutó una depuración exhaustiva (limpieza de valores duplicados, normalización de formatos de texto y numéricos, y tratamiento de campos vacíos). Esta etapa concluyó con un proceso de validación formal que garantizó que los datos consolidados fueran consistentes y reflejaran la realidad institucional.

Análisis de Patrones e Indicadores Críticos (Modelado conceptual y analítico)

Con el conjunto de datos validado, se realizaron análisis estadísticos estructurados para identificar las variables que mayor impacto o correlación tienen sobre el desempeño académico de los estudiantes. Aquí se generaron las primeras visualizaciones por facultades, programas académicos y periodos, transformando las filas de datos en tendencias claras y perfiles estudiantiles comprensibles.

Evaluación de Viabilidad Predictiva y Estándares (Evaluación de viabilidad y calidad)

En lugar de implementar directamente un modelo algorítmico a ciegas, se ejecutó un estudio riguroso para auditar si los datos históricos de los años 2018 a 2023 contaban con el volumen, equilibrio estadístico y cumplimiento de normativas legales (como la protección de datos personales y Habeas Data) requeridos por el ICFES. Como resultado, se documentaron conclusiones estratégicas detallando qué programas cuentan con la madurez de datos óptima y cuáles requieren acciones previas.

Estructuración y Construcción del Dashboard Web (Implementación)

Una vez definidos los hallazgos explicativos y las variables esenciales, se diseñó la arquitectura visual y funcional (*mockup*) del tablero de control. Posteriormente, se llevó a cabo el desarrollo técnico del Dashboard interactivo, traduciendo los requerimientos en un entorno dinámico basado en filtros dinámicos por cohorte, programa e indicadores clave de rendimiento (KPIs).

Aseguramiento, Pruebas y Despliegue (Retroalimentación continua)

El ciclo concluyó con una fase de control de calidad, donde se aplicaron pruebas funcionales para garantizar el correcto funcionamiento de los filtros, la velocidad de carga de las gráficas y la veracidad de los cálculos expuestos. Finalmente, se realizó el despliegue del aplicativo en el entorno institucional correspondiente,

poniéndolo a disposición de la Universidad como la herramienta oficial de consulta y diagnóstico.

Junto al modelo de desarrollo, también se definieron las actividades de la Pasantía, así como las condiciones técnicas necesarias para su implementación. Actividades que también se distribuyeron en tres grandes (3) fases, como podemos ver en la siguiente tabla.

Tabla 2. Actividades desarrolladas y condiciones técnicas.

Actividad Desarrollada	Condiciones Técnicas y Entorno Operativo
FASE I: GESTIÓN DE DATOS	
Recolección de datos desde la plataforma ICFES y fuentes institucionales.	• Espacio físico: Oficina asignada (Sitio de trabajo) y equipo de cómputo personal. • Conectividad: Conexión estable a internet para descarga de repositorios públicos. • Entorno de software: Plataforma ICFES interactiva, sistemas internos de la Universidad.

Limpieza e integración en un dataset unificado.	• Lenguaje y Librerías: Python, entorno Pandas para manipulación matricial de datos. • Herramientas ETL: Scripts estructurados para la extracción, transformación y unificación de registros. • Infraestructura: Cuadernos de trabajo (Jupyter Notebook) bajo entorno virtual Anaconda (Conda).
Validación de datos integrados.	• Herramientas analíticas: Dataframes de Pandas, funciones de verificación lógica de integridad referencial. • Criterios operativos: Auditoría de tipos de datos, consistencia transaccional y control de valores nulos.
FASE II: ANÁLISIS Y VIABILIDAD	
Análisis estadístico descriptivo e identificación de patrones y variables críticas.	• Librerías Core: NumPy, Pandas (módulos estadísticos para medias, desviaciones y correlaciones).

	• Entorno técnico: Jupyter Notebook para la documentación inmediata del análisis exploratorio.
Visualización preliminar de datos por programa y cohorte.	• Librerías gráficas: Matplotlib y Seaborn en Python para la creación de gráficos de dispersión, diagramas de caja e histogramas. • Enfoque técnico: Segmentación lógica de datos mediante agrupamientos indexados.
Documentación de hallazgos iniciales como insumo para el dashboard.	• Software de reporte: Herramientas ofimáticas y Markdown para consolidación de indicadores clave de rendimiento (KPIs).
Revisión de estándares ICFES para uso de datos en modelos predictivos.	• Fuentes consultadas: Documentación técnica oficial del ICFES, normativas legales vigentes de Habeas Data y protección de datos.
Estudio de viabilidad técnica por programa académico con datos 2018–2023.	• Enfoque analítico: Evaluación cualitativa y cuantitativa de la densidad y representatividad de las cohortes 2018-2023.

	• Instrumento: Matrices de evaluación técnica cruzada y diagnóstico de volumen crítico de datos.
Documentación de conclusiones y recomendaciones sobre viabilidad del modelo.	• Entregable técnico: Formatos estructurados de informe ejecutivo orientados a la toma de decisiones institucionales.
FASE III: DESARROLLO Y DESPLIEGUE	
Diseño de mockup y arquitectura del dashboard.	• Software de diseño: Bocetos estructurales (<i>wireframes</i>) y planeación lógica de los flujos de navegación del usuario. • Arquitectura: Modelado de la estructura de tablas de datos relacionales óptimas para el motor visual.
Desarrollo del dashboard web.	• Ecosistema principal: PowerBI Desktop y entorno PowerBI Web para la publicación del tablero.

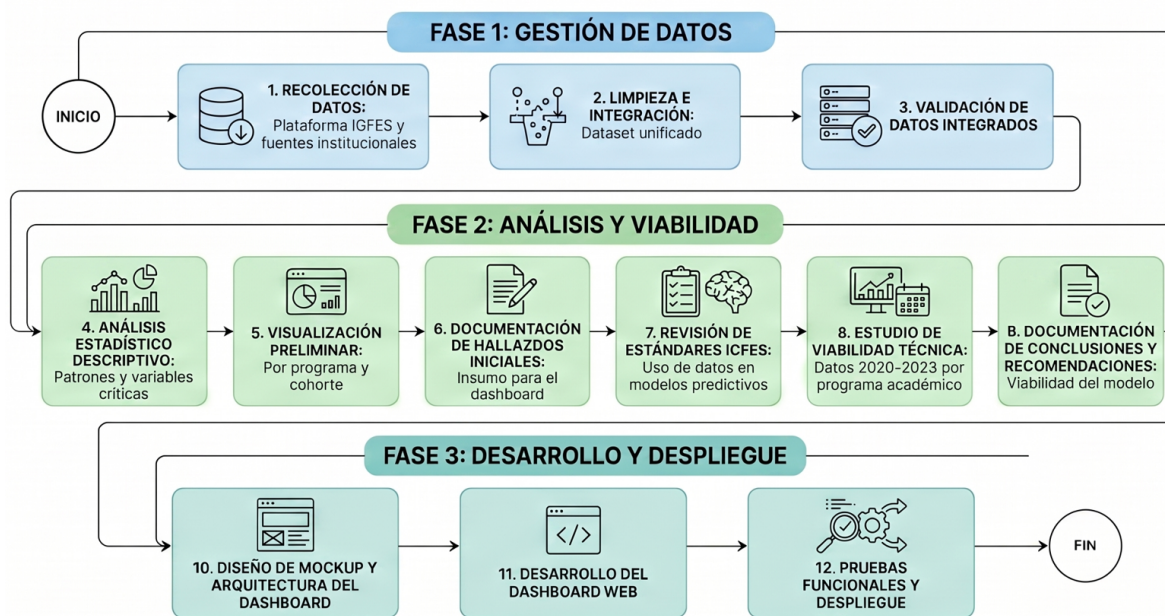
	• Modelado interno: Lenguaje DAX para el cálculo dinámico de métricas, promedios y tabulaciones cruzadas.
Pruebas funcionales y despliegue.	• Entorno de despliegue: Servidor o espacio de trabajo en la nube de la institución (PowerBI Service). • Metodología de prueba: Auditoría manual de filtros dinámicos, validación cruzada de valores contra el dataset unificado y pruebas de experiencia de usuario.

Fuente: Autor

Si bien la fase preliminar de exploración y análisis de los datos (Tabla 2) contempló originalmente la generación de gráficos estáticos mediante las librerías *Matplotlib* y *Seaborn* en Python, en consonancia con el segundo objetivo específico, se determinó metodológicamente adecuar dichas salidas analíticas concluyentes de la fase preliminar de exploración, hacia las vistas interactivas nativas en Power BI (reflejadas en las Figuras 17 a 20). Esta transición no sólo robustece el análisis exploratorio al permitir la identificación dinámica de brechas y patrones por programa y cohorte, sino que apalanca así el cumplimiento del cuarto objetivo específico del proyecto.

En la siguiente figura (Figura 3), se muestra el diagrama de actividades que se utilizó en el transcurso de la Pasantía.

Figura 3 Diagrama de actividades y secuencia de flujo.



Fuente: Autor

DESARROLLO DEL PROYECTO

El desarrollo de la pasantía se ejecutó bajo un enfoque metodológico iterativo, estructurando el ciclo de vida del proyecto desde la adquisición de datos en bruto hasta el despliegue de una solución integral. Partiendo de las fases, subfases, hasta las actividades realizadas de la Pasantía, se desarrollaron los entregables correspondientes, tal y como se muestra en la siguiente tabla:

Tabla 3. Fechas de actividades desarrolladas y productos entregados.

Fase	Actividad Desarrollada	Fechas	Productos Entregados
FASE I: GESTIÓN DE DATOS	Diagnóstico de fuentes de datos, recolección de datos desde la plataforma ICFES y fuentes institucionales.	21/10/2024 - 25/10/2024	Guías, diccionarios de datos y bases de datos descargadas.
FASE I: GESTIÓN DE DATOS	Limpieza e integración en un dataset unificado.	28/10/2024 - 15/11/2024	Código de limpieza para datos históricos.
FASE I: GESTIÓN DE DATOS	Validación de datos integrados.	18/11/2024 - 29/11/2024	Código de integración y validación de cruce para datos históricos. Código unificado PDFs Saber Pro.

FASE II: ANÁLISIS Y VIABILIDAD	Análisis estadístico descriptivo e identificación de patrones y variables críticas.	02/12/2024 - 27/12/2024	Identificación de variables críticas y complementarias.
FASE II: ANÁLISIS Y VIABILIDAD	Visualización preliminar de datos por programa y cohorte.	02/12/2024 - 27/12/2024	Visualización preliminar de los hallazgos iniciales.
FASE II: ANÁLISIS Y VIABILIDAD	Documentación de hallazgos iniciales como insumo para el dashboard.	23/12/2024 - 27/12/2024	Documentación de hallazgos iniciales.
FASE II: ANÁLISIS Y VIABILIDAD	Revisión de estándares ICFES para uso de datos en modelos predictivos.	30/12/2024 - 03/01/2025	Notas de viabilidad técnica.
FASE II: ANÁLISIS Y VIABILIDAD	Estudio de viabilidad técnica por programa académico con datos 2018–2023.	30/12/2024 - 10/01/2025	Matrices de evaluación por programa y cohorte académico.
FASE II: ANÁLISIS Y VIABILIDAD	Documentación de conclusiones y recomendaciones sobre viabilidad del modelo.	30/12/2024 - 10/01/2025	Documentación de conclusiones y recomendaciones

			sobre viabilidad del modelo.
FASE III: DESARROLLO Y DESPLIEGUE	Diseño de mockup y arquitectura del dashboard.	13/01/2025 - 24/01/2025	Diseño de Mockup preliminar.
FASE III: DESARROLLO Y DESPLIEGUE	Desarrollo del dashboard web.	13/01/2025 - 24/01/2025	Archivo (.pbix) del aplicativo de datos históricos Saber Pro y Saber 11. Archivo (.pbix) del aplicativo de datos PDFs con resultados individuales.
FASE III: DESARROLLO Y DESPLIEGUE	Pruebas funcionales y despliegue.	27/01/2025 - 07/02/2025	Manual de usuario del Código y aplicativos, y enlaces de los despliegues.

Fuente: Autor

Como se evidencia en la tabla, antes de comenzar la fase técnica, se realizó una investigación diagnóstica para localizar y comprender la estructura de los datos crudos. Se consultaron manuales técnicos y fuentes oficiales para interpretar la semántica de las variables. En colaboración con el director del proyecto, se evaluaron fuentes internas (Universidad) y externas (ICFES, CNA), identificando la relevancia de integrar los resultados de las pruebas Saber 11 como un factor de valor agregado para futuros análisis comparativos.

Recursos utilizados:

- Guía para la interpretación del Reporte de resultados en SABER PRO (Fuente: CNA)
- Resultados individuales pruebas Saber Pro (Universidad Santo Tomás - Seccional Tunja)
- Informe técnico: Medición de los efectos de la educación superior en Colombia sobre el aprendizaje estudiantil (ICFES)
- Reporte de resultados por aplicación examen saber pro - instituciones de educación superior (Universidad Santo Tomás/ICFES)
- Compendio de niveles de desempeño, pruebas saber Pro y pruebas saber 11 (ICFES)
- Diccionario de datos, pruebas saber Pro y pruebas saber 11 (ICFES)
- Resultados anonimizados pruebas saber Pro y pruebas saber 11. (ICFES)
- Llave_saber11_saberpro (ICFES)

Adicionalmente, para garantizar la correcta ejecución de las fases de gestión de datos, análisis de viabilidad y el despliegue del aplicativo, se definió un ecosistema tecnológico con las herramientas a utilizar. En la tabla siguiente *[Tabla 4]*, se detalla la relación del software, los lenguajes y librerías, junto con sus versiones y el propósito específico que cumplieron dentro de la pasantía.

Tabla 4. Software, los lenguajes y librerías utilizadas

Herramienta / Tecnología	Versión Sugerida / Utilizada	Propósito / Uso en el Proyecto
Python	3.10+	Lenguaje de programación principal empleado para la estandarización, limpieza, integración y validación del dataset unificado de las pruebas Saber 11 y Saber Pro.
Pandas	2.1+	Librería de Python utilizada para la manipulación matricial de datos (Dataframes), procesos ETL (extracción y transformación) y auditoría de tipos de datos y valores nulos.
NumPy	1.26+	Librería complementaria de Python usada para el cálculo de módulos estadísticos (medias, desviaciones y correlaciones) durante el análisis estadístico descriptivo.
Matplotlib y Seaborn	3.8+ / 0.13+	Librerías gráficas implementadas para la visualización preliminar de datos, creación de gráficos de dispersión, diagramas de caja e histogramas por programa y cohorte.

Jupyter Notebook (vía Anaconda)	Jupyter 4.1+	Infraestructura de entorno virtual y cuadernos de trabajo interactivos empleada para la documentación inmediata del análisis exploratorio y ejecución de scripts.
Power BI Desktop	2.13+	Ecosistema principal para el diseño de la arquitectura del dashboard, modelado interno de la estructura de tablas relacionales y creación de visualizaciones interactivas.
Power BI Service	N/A (Cloud)	Entorno de servidor o espacio de trabajo en la nube de la institución utilizado para el despliegue final, auditoría de filtros dinámicos y pruebas funcionales.
DAX (Data Analysis Expressions)	N/A (Integrado en PBI)	Lenguaje de expresiones de análisis de datos utilizado dentro de PowerBI para el cálculo dinámico de métricas, promedios y tabulaciones cruzadas.
Sharepoint	Microsoft 365	Directorio cloud, contenido en el paquete de Microsoft 365 de la Universidad, utilizado para la centralización de datos.

Fuente: Autor

Fase 1. Gestión de datos

En el transcurso de la pasantía se utilizaron dos fuentes de datos; los datos históricos públicos (Saber Pro y Saber 11) compartidos por la entidad ICFES, y los resultados individuales más recientes (Archivos .PDF) de las pruebas Saber Pro proporcionados por la Universidad. Fuentes que requirieron procesos de Gestión de datos particulares. Por consiguiente, para evitar confusión en esta 'Fase 1. Gestión de datos', los procesos de las dos fuentes de datos se van a denominar como 'Datos históricos públicos' y 'Datos recientes internos' respectivamente.

Recolección de datos

→ Para datos históricos públicos:

A continuación, se documenta mediante las Figuras 4, 5, 6 y 7 el procedimiento técnico implementado para la obtención de microdatos desde Datalcfes, que es un directorio en Sharepoint compartido por la entidad ICFES, cuyo análisis metodológico se detalla al finalizar el procedimiento.

Procedimiento:

1. Ingresar al portal oficial de datos del Icfes (Datalcfes).
2. Completar el Formulario de registro. Este paso es indispensable para declarar que el uso de la información tiene fines académicos o de investigación y para aceptar las políticas de tratamiento de datos.

Figura 4 Imagen tomada del Formulario DataIcfes.

The image shows a web form titled "Formulario de registro - DataIcfes". At the top left is the "DataIcfes" logo. Below the title, there is a paragraph of text explaining that the information collected is confidential and will only be used for the purposes stated, in accordance with the principle of confidentiality established in the Law of Habeas Data. It states that these data will not be used for any other purpose different from the one mentioned, and thanks the user for completing the survey in full, clear, and sincere. Below this, there is a section for authorization, stating that the user authorizes Icfes to treat their personal data registered here, in accordance with the Icfes Policy on the Treatment of Personal Information, for the purposes described. It provides a link to the policy: <https://www.icfes.gov.co/ley-1581-de-2012-proteccion-de-datos-personales>. At the bottom, it asks the user to answer the following questions before registering on Data Icfes.

Fuente: DataIcfes

En la Figura 4 se observa un formulario de inscripción en formato de encuesta, etapa inicial indispensable, ya que define el origen y la procedencia legal de la información, previniendo la extracción de fuentes alteradas o no institucionales.

3. Una vez enviado el formulario, al correo registrado va a llegar un enlace que les va a direccionar a un documento compartido. *[Figura 5]*

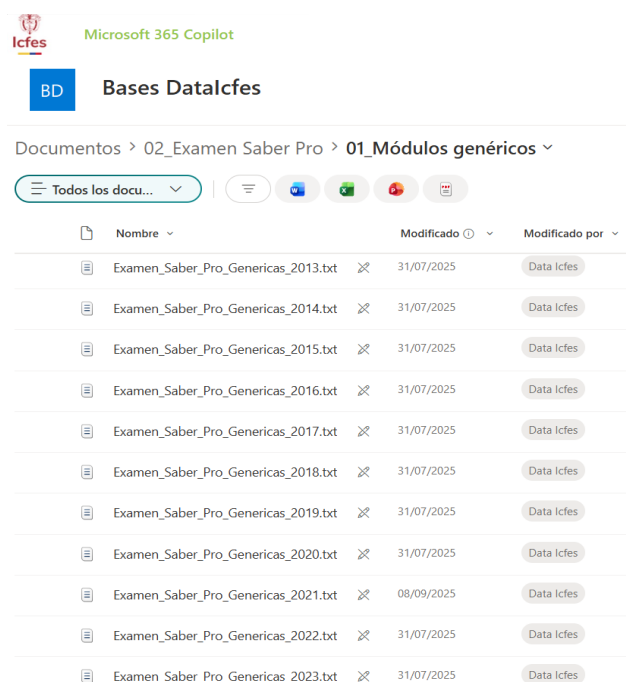
Figura 5 Imagen tomada del Documento DataIcfes



Fuente: Datalcfes

4. En el documento, ubicar el índice y seleccionar la prueba específica de interés: Saber 11 (para educación media) o Saber Pro (para educación superior).
5. Al seleccionar el título de la prueba, nos va a redireccionar al directorio Datalcfes en Sharepoint. Ahí podemos descargar los archivos de los resultados soportados generalmente en formatos de texto. (.txt).

Figura 6 Imagen tomada en Sharepoint, directorio Datalcfes



Nombre	Modificado	Modificado por
Examen_Saber_Pro_Genericas_2013.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2014.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2015.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2016.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2017.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2018.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2019.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2020.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2021.txt	08/09/2025	Data Icfes
Examen_Saber_Pro_Genericas_2022.txt	31/07/2025	Data Icfes
Examen_Saber_Pro_Genericas_2023.txt	31/07/2025	Data Icfes

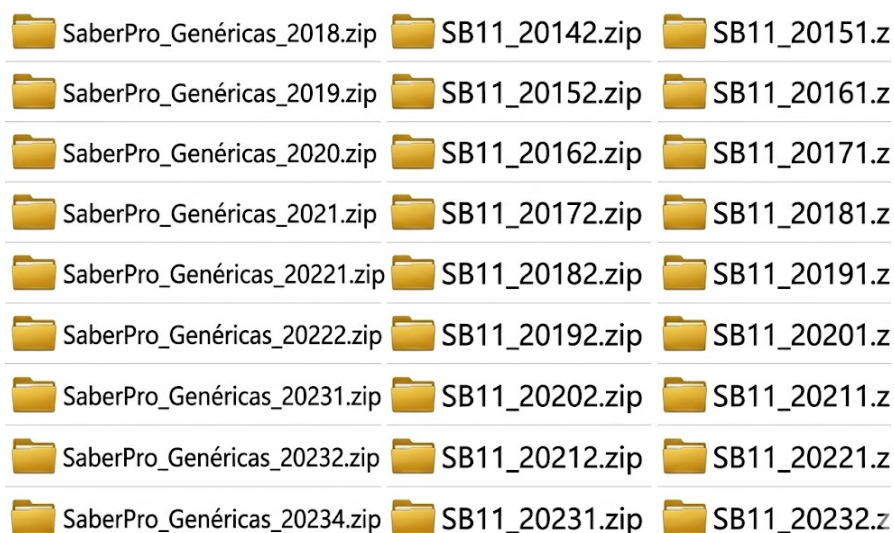
Fuente: Datalcfes

La Figura 6 muestra los microdatos disponibles. La descarga de los microdatos de las pruebas SaberPro y pruebas Saber11 se validó mediante la selección estricta del periodo temporal bajo estudio, garantizando que el volumen de registros inicial coincida de manera uniforme con las bases de datos de referencia académica.

6. Es estrictamente necesario descargar el "Diccionario de variables" o "Diccionario de datos". Estos documentos son la guía que permite entender qué significa cada columna dentro de los microdatos para poder realizar el análisis.

En la siguiente Figura, se muestra la evidencia de la extracción de datos históricos correspondientes a las pruebas Saber 11 (periodos 2014 a 2023-2) y Saber Pro (periodos 2018 a 2023-2) desde la fuente oficial (ICFES).

Figura 7 Imagen tomada de los archivos descargados de DataIcfes (.zip)



Fuente: Autor

Nota: Se prepararon veinticinco (25) archivos. [Figura 7]

→ Para datos recientes internos:

Los resultados individuales proporcionados por la Universidad en formato PDF se obtuvieron en un proceso progresivo, a lo largo de las primeras etapas. Por consiguiente, se empezó un proceso de reconocimiento sobre los documentos PDF disponibles, para identificar variables, y poder compararlo paralelamente con el avance de los datos históricos públicos.

Nota: En total se obtuvieron 1212 documentos (PDF), correspondiente a los resultados individuales de las pruebas Saber Pro para el periodo 2023-2024, con un promedio de ocho (8) páginas por PDF.

Limpieza e integración

Esta fase fue fundamental para establecer la integridad del proceso. Se aplicaron técnicas de ingeniería de datos para transformar la información dispersa en un conjunto de datos estructurado, enfocándose en la creación de una fuente de datos confiable.

→ Para datos históricos públicos:

Utilizando el entorno de desarrollo Jupyter, se programó un script en Python para normalizar formatos, limpieza y Validación; teniendo en cuenta que los microdatos provenían de archivos de texto (.txt). Esto incluyó la unificación de separadores (“;”, “,” , “,” , “↵”, “|”), el tratamiento de valores nulos, la estandarización de mayúsculas/minúsculas, corrección de tipos de datos anómalos, valores corridos, rangos de valores inconsistentes y demás.

Las siguientes figuras muestran parte de este proceso; desde la verificación de los separadores, pasando a la normalización de formatos, hasta la estandarización de valores.

Figura 8 verificación de separadores en archivos .txt

```
# Función para verificar los separadores en el archivo
def check_file_separators(filepath):
    with open(filepath, 'r', encoding='utf-8') as file:
        first_line = file.readline().strip()
        separadores = [';', ',', '-', '\n']

        for sep in separadores:
            if sep in first_line:
                print(f"El separador '{sep}' está presente.")
                return sep # Retorna el separador encontrado
        print("No se encontró ningún separador estándar.")
        return None # Retorna None si no se encuentra un separador

separadores = check_file_separators(filepath)
```

Fuente: Autor

Figura 9 Normalización de formatos

```
columns = ['MOD_RAZONA_CUANTITAT_PUNT', 'MOD_RAZONA_CUANTITAT_DESEM', 'MOD_RAZONA_CUA',
           'MOD_COMPETEN_CIUADADA_PNBC', 'MOD_COMUNI_ESCRITA_PUNT', 'MOD_COMUNI_ESCRITA_']

data_types = {'MOD_RAZONA_CUANTITAT_PUNT': int, 'MOD_RAZONA_CUANTITAT_DESEM': int, 'MOD_LECTURA_CRITICA_PNBC': int, 'MOD_COMPETEN_CIUADADA_PUNT': int, 'MOD_COMUNI_ESCRITA_PNBC': int, 'MOD_COMUNI_ESCRITA_PNAL': int, 'MOD_COMUNI_ESCRITA_PNBC': int, 'MOD_IN'}

ranges = {'MOD_RAZONA_CUANTITAT_PUNT': (0, 300), 'MOD_RAZONA_CUANTITAT_DESEM': (1,4), 'MOD_LECTURA_CRITICA_PNBC': (1, 100), 'MOD_COMPETEN_CIUADADA_PUNT': (0,300), 'MOD_COMUNI_ESCRITA_PNAL': (1, 100), 'MOD_COMUNI_ESCRITA_PNBC': (1, 100), 'MOD_IN'}

categorical_defaults = {'MOD_INGLES_DESEM': '-A1'}
```

Fuente: Autor

Figura 10 Estandarización de valores vacíos/nulos.

```
empty_values = [None, "", " ", "\t", "\n", "\r", np.nan]

def convert_and_validate(df, columns, data_types, ranges, categorical_defaults):
    errors = []

    for col in columns:
        # Verificar si la columna existe en el DataFrame
        if col not in df.columns:
            print(f"Columna {col} no encontrada. Se ignora.")
            continue

        print(f"Procesando columna: {col}") # Diagnóstico

        # Verificar si es una columna categórica con valor predeterminado
        if col in categorical_defaults:
            print("Valores originales:", df[col].unique()) # Diagnóstico

            # Reemplaza los valores vacíos en la columna con el valor predeterminado
            df[col] = df[col].replace(empty_values, np.nan).fillna(categorical_defaults[col])
            print("Después de rellenar valores vacíos:", df[col].unique()) # Diagnóstico
        else:
```

Fuente: Autor

Posteriormente, se desarrolló un módulo de validación en el Script de Jupyter, diseñado para identificar y filtrar de manera precisa a los estudiantes vinculados a la Universidad Santo Tomás - Seccional Tunja dentro de los microdatos nacionales del ICFES desde el año 2018 hasta el año 2023-2. Con el fin de garantizar que el análisis se centrará exclusivamente en la población objetivo de la pasantía.

Figura 11 Estudiantes que presentaron su examen saber Pro en el año 2018

```
[27] import pandas as pd

# Lee el archivo de texto con separador unificado
sbpro = pd.read_csv('./SP_2018.txt', sep=',')

sbpro.columns = sbpro.columns.str.upper()

[28] # Filtrar por institución
filtro = (sbpro['INST_NOMBRE_INSTITUCION'] == 'UNIVERSIDAD SANTO TOMAS-TUNJA')

merged_df = sbpro[filtro]

[29] conteo_por_nucleo = merged_df.groupby('ESTU_NUCLEO_PREGRADO').size()
conteo_por_nucleo
```

...	ESTU_NUCLEO_PREGRADO	
	ADMINISTRACIÓN	107
	ARQUITECTURA	100
	CONTADURÍA PÚBLICA	68
	DERECHO Y AFINES	159
	INGENIERÍA CIVIL Y AFINES	180
	INGENIERÍA DE SISTEMAS, TELEMÁTICA Y AFINES	14
	INGENIERÍA ELECTRÓNICA, TELECOMUNICACIONES Y AFINES	17
	dtype: int64	

Fuente: Autor

Como podemos ver en la figura anterior, para el año 2018, los programas académicos 'Derecho y afines' e 'Ingeniería civil y afines' tuvieron la mayor cantidad de estudiantes presentando las pruebas saber Pro, por el contrario, los programas de 'Ingeniería de Sistemas' e 'Ingeniería electrónica' fueron los de menor presencia en estas pruebas.

→ Para datos recientes internos:

Con los datos obtenidos hasta el momento por parte de la Universidad, se hizo un análisis de las opciones sugeridas para el procesamiento de los archivos PDFs y la extracción de los datos pertinentes:

- Reconocimiento Óptico de Caracteres (OCR)
 - Se descartó esta opción por complejidad y precisión innecesaria cuando los PDFs ya están en texto y no imagen.
- Modelos Basados en IA
 - Se descartó esta opción por dependencia de versiones, complejidad en la configuración y la imposibilidad de un control a nivel infra.
- Expresiones Regulares (RE)
 - Se eligió esta opción por la simplicidad y eficiencia en la extracción de patrones específicos, control total sobre el proceso y la facilidad de mantenimiento.

Después de haber elegido la forma de extracción más viable, utilizando el entorno de desarrollo Jupyter, se programó un script en Python para la extracción, normalización, definición de formatos, limpieza y validación. Se identificaron los datos o variables a extraer, que llegaron a ser dieciocho (18). No hubo exclusiones particulares como en los datos históricos públicos, a excepción de las 'competencias específicas' por programa académico, ya que no se tienen pensado utilizar en este proceso.

En la Figura 12 se muestra el uso de Expresiones Regulares (RE), opción que aprovecha la estructura común entre documentos.

Figura 12. Imagen recortada del uso de expresiones regulares en la extracción de los datos.

```
import os
import pdfplumber
import pandas as pd
from unicode import unicode
from tqdm import tqdm
import re

# Ruta a la carpeta con los PDFs
folder_path = "./fuentes/SaberPro"

# Creación de una lista para almacenar los resultados de cada PDF
all_data = []

# Patrones RE
numero_cedula_pattern = r"(C\.C\.|T\.|I\.)\s*(\d+)\s*EK"
numero_registro_pattern = r"\bEK[0-9A-Za-z]+\b"
periodo_pattern = r"EK(\d{4})"
```

Fuente: Autor

Al terminar de identificar el patrón para las dieciocho (18) variables, se dio inicio a la extracción de las mismas, como se ve en la siguiente Figura.

Figura 13. Tiempo promedio de análisis por PDFs (1212): 1.30s por archivo.

Procesando PDFs: 100% | 1212/1212 [26:11<00:00, 1.30s/it]
Extracción completada.

Fuente: Autor

Como se puede apreciar en la Figura 13, se procesaron 1212 documentos PDF, extrayendo individualmente dieciocho (18) variables dentro de ocho (8) páginas en promedio. Un proceso que duró veintiséis minutos y once segundos, con un promedio de un segundo y treinta milisegundos por documento.

Validación de datos integrados

→ Para datos históricos públicos:

Se implementó un código de integración que ejecutó el cruce de variables (Data Merging) entre los registros validados de Saber Pro y el histórico de las pruebas Saber 11. Para ello, se utilizaron las llaves únicas de identificación proporcionadas por la entidad ICFES para el cruce de información, resolviendo las discrepancias de formato entre distintos periodos anuales.

Figura 14 Cruce Saber Pro con Saber 11.

Cruce

```
import pandas as pd

# Merge inicial usando la columna ESTU_CONSECUTIVO_SB11 de la llave
merged_df = pd.merge(sb11, llave, left_on='ESTU_CONSECUTIVO', right_on='ESTU_CONSECUTIVO_SB11')

# Merge final usando la columna ESTU_CONSECUTIVO_SBPRO de la llave
merged_df = pd.merge(merged_df, sbpro, left_on='ESTU_CONSECUTIVO_SBPRO', right_on='ESTU_CONSECUTIVO')

# Eliminar columnas redundantes (si es necesario)
merged_df = merged_df.drop(['ESTU_CONSECUTIVO_SB11', 'ESTU_CONSECUTIVO_SBPRO'], axis=1)

merged_df.to_csv("raw_cruce2018.txt", sep=',', index=False)
```

Fuente: Autor

La Figura 14 nos muestra un poco sobre cómo se realizó la integración entre los resultados Saber Pro y Saber 11. Como resultado de la integración, se pudo calcular la tasa de éxito de los cruces segmentada por carrera o Núcleo Básico de Conocimiento (NBC). Esta métrica fue fundamental para el estudio de viabilidad técnica en la aplicación de modelos predictivos, porque consiguió determinar la significancia estadística por programa académico.

Figura 15 Resultados del cruce, distribuido por NBC.

```
# Filtrar por institución  
filtro = (rawc['INST_NOMBRE_INSTITUCION'] == 'UNIVERSIDAD SANTO TOMAS-TUNJA')  
  
merged_df_raw = rawc[filtro]
```

```
conteo_por_nucleo_raw = merged_df_raw.groupby('ESTU_NUCLEO_PREGRADO').size()  
conteo_por_nucleo_raw
```

```
ESTU_NUCLEO_PREGRADO  
ADMINISTRACIÓN      47  
CONTADURÍA PUBLICA   6  
INGENIERÍA CIVIL Y AFINES  4  
dtype: int64
```

Fuente: Autor

Durante esta etapa se presentó la primera gran dificultad técnica: se identificó una inconsistencia severa en las llaves de cruce entre los registros de Saber 11 y Saber Pro, provocada por cambios de formato a lo largo de los años por parte de la entidad ICFES.

★ *Entregables:*

Con el propósito de consolidar un conjunto de datos unificado y debidamente documentado que sirviese de base analítica para el proyecto, se procuró mantener la estructura que ya viene utilizando la entidad ICFES, que es, por consiguiente, su diccionario de datos. Las fuentes de datos originales (archivos en formato Microsoft Excel correspondientes a las bases históricas de Saber 11, Saber Pro histórico y Saber Pro reciente), así como los repositorios de código fuente desarrollados en Python, han sido centralizados y dispuestos formalmente para la auditoría y consulta institucional en el SharePoint Universitario, específicamente en el directorio de acceso restringido denominado Pasantía 2024.

→ Para datos recientes internos:

En el proceso de validación e integración se validaron los campos críticos como el Puntaje Global (y sus percentiles por NBC), los puntajes por competencia, niveles de desempeño, habilidades asociadas y percentiles específicos para grupos étnicos o condiciones de discapacidad.

De ello, se identificaron escenarios que podrían afectar el análisis e integración de los datos, tales como:

- Inconsistencias de Registro:
Estudiantes sin Programa Académico definido en el reporte o registros duplicados por doble titulación.
- Variabilidad Estructural:
Reportes con secciones adicionales de percentiles para estudiantes pertenecientes a grupos étnicos o con condiciones de discapacidad.
- Casos de Desempeño Atípico:
Identificación de niveles de desempeño en “0” (competencia no presentada) o nivel “1” (competencia presentada sin respuestas registradas).
- Recurrencia:
Casos de estudiantes que presentaron la prueba en años consecutivos (2023 y 2024).

Teniendo en cuenta lo anterior, se creó un Script en Jupyter que unifica el preprocesamiento, limpieza e integración de los 1212 archivos PDF, y de cualquier archivo futuro relacionado.

El alcance de esta implementación se definió bajo tres objetivos estratégicos:

- **Automatización completa del proceso:** Mediante el desarrollo de un script en Jupyter orientado a la preparación de los datos, se eliminó la intervención manual en la captura de datos no estructurados desde los archivos PDF.

- **Centralización de los datos:** Se logró consolidar la información dispersa en un único repositorio estructurado (Sharepoint), facilitando la trazabilidad y la integridad de los registros académicos.
- **Facilidad de uso para personal no técnico:** La solución se diseñó para que administrativos pudieran ejecutar el proceso y consultar los resultados sin requerir conocimientos profundos en programación o ingeniería de datos.

La ejecución de este flujo permitió la transformación de cientos de documentos PDF en un conjunto de datos tabular consolidado (Excel). Este hito no sólo resolvió un problema administrativo inmediato, sino que permitió la construcción de un almacén de datos estrictamente necesario para habilitar futuras fases de analítica avanzada y Machine Learning.

★ *Entregables:*

Se entregó un Script de uso único en Jupyter, Script para el preprocesamiento, limpieza e integración de los PDF.

Fase 2. Análisis y viabilidad

Esta fase del proyecto tuvo dos enfoques; la preparación de los datos para la creación del aplicativo Dashboard, y el estudio técnico de viabilidad para el uso de modelos predictivos.

El primer enfoque cumplió con dos objetivos claros centrado en los 'datos históricos públicos' y 'datos recientes internos':

- Identificar patrones y brechas de desempeño académico por programa y cohorte.
- Centralizar el acceso a los resultados Saber Pro y apoyar la toma de decisiones de la Dirección de Aseguramiento de la Calidad.

El segundo enfoque se centra en el estudio técnico de viabilidad para el uso de modelos predictivos. En este enfoque los 'Datos históricos públicos' tomaron relevancia debido al cruce exitoso entre los resultados Saber 11 y los resultados Saber Pro, cumpliendo así con el requisito básico para el estudio técnico de viabilidad en la aplicación de modelos predictivos.

Los 'Datos recientes internos' desafortunadamente no pudieron incluirse en el estudio por la falta del dato/llave identificador para el cruce de resultados con las pruebas Saber 11.

Análisis descriptivo y exploratorio

Este proceso se centró en la investigación y entendimiento de los datos, con el objetivo de direccionar los siguientes pasos del proceso.

Se utilizaron los siguientes recursos:

- Compendio de niveles de desempeño, para pruebas saber Pro y pruebas saber 11 (ICFES) (Ver Figura 16)
- Informe técnico: Medición de los efectos de la educación superior en Colombia sobre el aprendizaje estudiantil (ICFES)
- Reporte de resultados por aplicación examen saber pro - instituciones de educación superior (Universidad Santo Tomás/ICFES)
- Diccionario de datos, pruebas saber Pro y pruebas saber 11 (ICFES)
- Guía para la interpretación del Reporte de resultados en SABER PRO (CNA)

Como una pequeña muestra de los recursos utilizados, la Figura 16 nos presenta la explicación de los niveles de desempeño para el módulo de lectura crítica de las pruebas Saber Pro. Recurso público y disponible por parte de la entidad ICFES.

Figura 16. Explicación niveles de desempeño para el módulo de lectura crítica (Saber Pro)



Fuente: ICFES

A partir de esta investigación diagnóstica, se identificaron y clasificaron las variables en críticas, complementarias e inviables, apoyándose también en los resultados obtenidos de la fase preliminar de limpieza y normalización para mitigar inconsistencias graves, tales como registros nulos o datos anómalos. Debido a que la estructura de los microdatos del ICFES presenta una alta variabilidad entre períodos anuales, oscilando entre un mínimo de 70 y un máximo de 130 variables por archivo, se priorizaron aquellos atributos con consistencia histórica. Asimismo, se seleccionaron los campos que viabilizan el análisis comparativo, específicamente los indicadores de desempeño de la prueba Saber 11 que permiten establecer una correlación directa con los resultados de Saber Pro.

Como resultado de esta fase, se definió el periodo de congruencia entre formatos, apoyado en los diccionarios de datos de las dos pruebas Saber. También se identificaron las variables que tanto el ICFES como el CNA utilizan o toman en cuenta al momento de interpretar los resultados.

En la etapa de 'Documentación de hallazgos' se van a enlistar las variables aquí encontradas, pero en resumen:

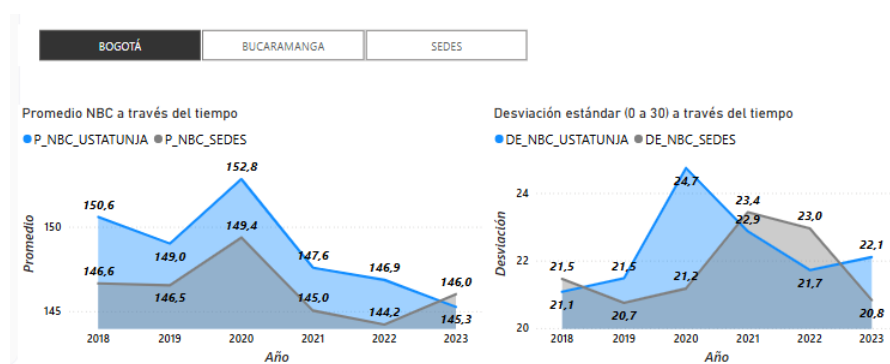
- Para Saber 11 se identificó:
Congruencia en variables y formatos entre los periodos 2014 a 2023-2
18 variables críticas.
7 variables complementarias.
- Para Saber Pro se identificó:
Congruencia en variables y formatos entre los periodos 2018 a 2023-2
24 variables críticas.
9 variables complementarias.

Visualización preliminar

Una vez estructurados los datos, se procedió a entender su comportamiento y distribución. Se prepararon visualizaciones preliminares para observar la distribución de los estudiantes por niveles de desempeño y grupos de referencia (Núcleo Básico de Conocimiento - NBC).

Las siguientes figuras muestran las evidencias de este proceso de análisis:

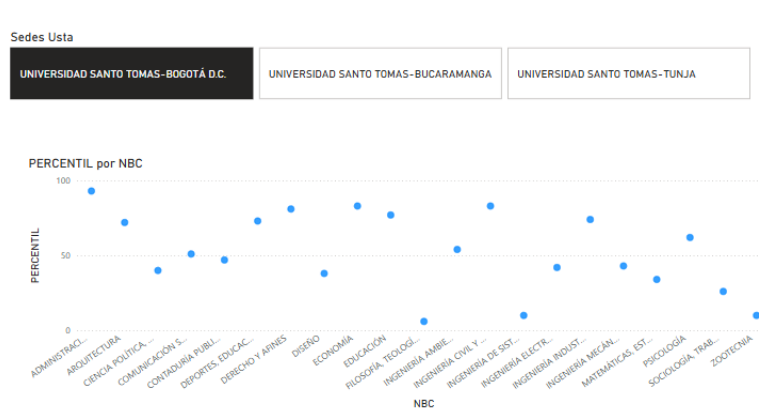
Figura 17 Sección NBC a través del tiempo (Filtro NBC y Sedes Usta)



Fuente: Autor

La Figura 17 forma parte de un proceso de cálculo de medidas de tendencia central y dispersión (promedios, modas, máximos, mínimos y desviaciones estándar) sobre los puntajes globales y por competencias.

Figura 18. Sección Percentil por NBC (Filtro Sedes Usta)



Fuente: Autor

Como podemos ver en la Figura 18, se previsualizan resultados del Percentil, en base al NBC, con una adición comparativa entre sedes.

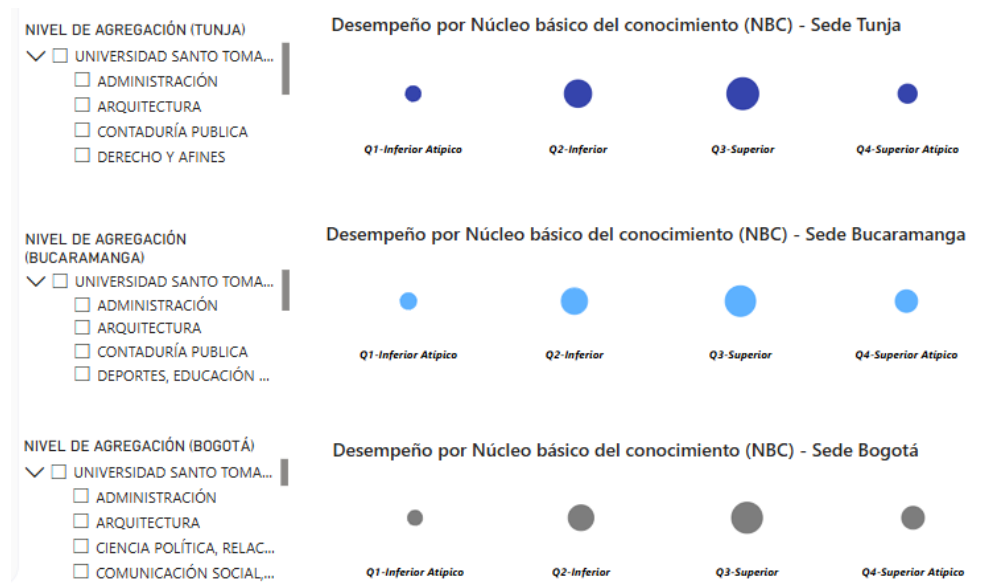
Figura 19 Sección desempeño por competencia (Filtro Nivel de agregación y Año)



Fuente: Autor

La Figura 19 nos previsualiza el desempeño por competencias generales, con posible filtrado de nivel de agregación y año. Y, como desglose a mayor profundidad, se creó un visual comparativo de mayor detalle, que podemos ver en la Figura 20.

Figura 20 Sección desempeño por NBC (Filtro Sedes Usta y Año)



Fuente: Autor

La Figura 20 presenta un indicador de Distribución de Estudiantes por Cuartiles de desempeño por NBC o programa académico, filtrado por sedes de la Universidad.

Documentación de hallazgos iniciales

Basado al estudio y visualización preliminar de los datos, se encontraron hallazgos muy importantes que incluyen métricas o variables críticas, tales como el Percentil Nacional, el Percentil por NBC y los Subniveles de Desempeño. Se documentaron los indicadores más relevantes para la toma de decisiones académicas por parte de la Universidad.

Como logro, se validaron los siguientes datos:

- Para los resultados globales:
 - Percentil por nivel de agregación. (Fuente: Reporte resultados Icfes).
 - Percentil por grupo de referencia (NBC) (Fuente: Reporte resultados Icfes).
 - Promedio del puntaje global y desviación estándar por nivel de agregación. (Fuente: Reporte resultados Icfes).
 - Promedio del puntaje global y desviación estándar por grupo de referencia (NBC) (Fuente: Reporte resultados Icfes).
- Para los resultados por carrera (NBC):
 - Promedio y desviación estándar por nivel de agregación. (Fuente: Reporte resultados Icfes).
 - Porcentaje de estudiantes por niveles de desempeño. (Fuente: Reporte resultados Icfes).
 - Promedio y desviación estándar por grupo de referencia. (Fuente: Reporte resultados Icfes).
 - Percentil por nivel de agregación. (Fuente: Reporte resultados Icfes).
 - Percentil por grupo de referencia. (Fuente: Reporte resultados Icfes).
 - Percentil por grupo de referencia (NBC). (Fuente: Reporte resultados Icfes).
 - Porcentaje de estudiantes por niveles de desempeño y grupo de referencia. (Fuente: Reporte resultados Icfes).
 - Promedio y desviación estándar por grupo de referencia (NBC). (Fuente: Reporte resultados Icfes).

Porcentaje de estudiantes por niveles de desempeño y grupo de referencia (NBC).
(Fuente: Reporte resultados Icfes).

- Para los resultados por módulo:

Promedio y desviación estándar por nivel de agregación. (Fuente: Reporte resultados Icfes).

Porcentaje de estudiantes por niveles de desempeño. (Fuente: Reporte resultados Icfes).

Promedio y desviación estándar por grupo de referencia. (Fuente: Reporte resultados Icfes)

Percentil por nivel de agregación. (Fuente: Reporte resultados Icfes).

Percentil por grupo de referencia. (Fuente: Reporte resultados Icfes).

Percentil por grupo de referencia (NBC). (Fuente: Reporte resultados Icfes).

Porcentaje de estudiantes por niveles de desempeño y grupo de referencia.
(Fuente: Reporte resultados Icfes).

Promedio y desviación estándar por grupo de referencia (NBC). (Fuente: Reporte resultados Icfes).

Porcentaje de estudiantes por niveles de desempeño y grupo de referencia (NBC).
(Fuente: Reporte resultados Icfes).

Porcentaje promedio de respuestas incorrectas asociadas a cada afirmación para la sede y cada grupo de referencia (NBC):

Si el porcentaje promedio de respuestas incorrectas es menor al 20% se asigna el color verde.

Si el porcentaje promedio de respuestas incorrectas es mayor o igual al 20% y menor

al 40% se asigna el color amarillo.

Si el porcentaje promedio de respuestas incorrectas es mayor o igual al 40% y menor

al 70% se asigna el color naranja.

Si el porcentaje promedio de respuestas incorrectas es mayor o igual al 70% se asigna el color rojo.

Revisión de estándares ICFES

En esta etapa se inició el estudio de viabilidad en la aplicación de modelos predictivos, que se basó, siguiendo la normatividad, en la documentación proporcionada por la entidad ICFES para la interpretación de los resultados, principalmente los del apartado de aporte relativo.

Citando la documentación de la entidad se definieron los siguientes requisitos iniciales:

- **Supuestos**
 - Manipulabilidad: Teóricamente, los estudiantes podrían estar expuestos a cualquier tratamiento.
 - No interferencia entre unidades: El resultado de un estudiante depende solamente de su asignación a un determinado tratamiento.
 - El supuesto métrico: Los resultados del puntaje de la prueba están en una escala de intervalo.
 - Homogeneidad: El efecto causal no varía como función de una característica del estudiante.
 - Tratamiento altamente ignorable: La asignación al tratamiento es esencialmente aleatoria, después de condicionarse por las variables de control.
 - Forma funcional: la forma funcional empleada para el control de las características del estudiante es la correcta.
- **Sesgos**
 - Sesgo en los coeficientes de los puntajes previos, debidos al error de medición.
 - Sesgo en las estimaciones del efecto escolar para IES con características contextuales favorables, efectos tipo B.
 - Sesgo en las estimaciones del efecto escolar para IES con características contextuales desfavorables.

- Requerimientos:
 - Para el cruce de las dos tablas (Saber 11 y Saber Pro): Se tienen en cuenta todos los individuos que presentaron el Examen Saber 11° entre 4 y 8 años antes de presentar el examen Saber Pro. Se limita el tiempo entre los dos exámenes para evitar atribuir al valor agregado de las instituciones experiencias externas de los estudiantes que no pueden ser observadas.
 - Para sesgos de puntaje: Se excluyen del estudio los registros que tienen dos o más módulos de Saber Pro con un puntaje igual a 0 o sin calificación, ya que son personas que por alguna razón no respondieron la prueba y pueden sesgar las estimaciones.
 - Requisitos para estimación del modelo de valor agregado: Instituciones dentro de un Núcleo Básico del Conocimiento (NBC), en las que existan al menos 25 estudiantes que hayan presentado Saber Pro, cuenten con su respectivo cruce en Saber 11° y representen el 40% o más del total de la población de la institución en el NBC correspondiente. Solo se hacen las estimaciones de valor agregado en aquellos NBC que cuenten con al menos cinco instituciones y en los que se cumplan las situaciones anteriores para los dos años considerados en el reporte (ej. 2017-2018)

Fuentes: Informe técnico, Icfes

Estudio de viabilidad técnica

En este punto se hizo la preparación para el estudio de viabilidad técnica, acompañado de los resultados obtenidos hasta este punto, que son:

- Resultados de las pruebas Saber 11 de estudiantes de la Universidad Santo Tomás-Seccional Tunja, periodos que van del 2014 a 2023.
- Resultados de las pruebas Saber Pro de estudiantes de la Universidad Santo Tomás-Seccional Tunja, Periodos que van del 2018 a 2023-2
- Cruce de resultados Saber 11 con Saber Pro, distribuido por año y programa académico.

A partir de estos registros históricos obtenidos en las fases anteriores (Preprocesamiento de datos y análisis exploratorio) se realizó un estudio de viabilidad técnica para determinar la factibilidad de implementar modelos predictivos. Se construyó una matriz de evaluación estratificada por Núcleo Básico del Conocimiento (NBC) o programa académico, fundamentada en los siguientes indicadores y criterios de calidad:

- **Población por NBC (T):** Volumen de estudiantes que presentaron la prueba en cada programa académico.
- **Criterio de Significancia Estadística ($T > 25$):** Se estableció como requisito mínimo una muestra superior a 25 exámenes por año y carrera para garantizar la fiabilidad de las estimaciones.
- **Porcentaje de Cruce Exitoso (C):** Proporción de registros Saber Pro que pudieron vincularse efectivamente con los resultados históricos de Saber 11.
- **Criterio de Representatividad ($C > 40\%$):** Se definió que el cruce de datos debía ser superior al 40% de la población del NBC para asegurar que el modelo refleje la tendencia real de la carrera.

Adicionalmente, el análisis incluyó la identificación de valores atípicos, tales como estudiantes con puntajes nulos (0) en uno o más módulos, cuya presencia fue evaluada para evitar sesgos en el entrenamiento de posibles algoritmos.

La Figura 21 nos presenta una muestra puntual del año 2022, con la identificación de los valores atípicos y los resultados del estudio de viabilidad para ese periodo.

Figura 21. Matriz de evaluación estratificada por Núcleo Básico del Conocimiento (NBC)

Saber Pro 2022:

Cruces exitosos: 669 (De 778 datos) [86%]

Estudiantes con puntaje de '0' en más de un módulo: 4

Estudiantes con puntaje de '0' en un módulo: 61 (Todos del módulo de comunicación escrita)

<i>NBC</i>	<i>TOTAL POR NBC</i>	<i>CRITERIO (T > 25)</i>	<i>CRUCE POR NBC</i>	<i>CRITERIO (C>40%)</i>
Administración	117	Sí	91	Sí
Arquitectura	189	Sí	168	Sí
Contaduría pública	25	Sí	17	Sí
Derecho y afines	139	Sí	123	Sí
Ingeniería civil y afines	96	Sí	85	Sí
Ingeniería de sistemas, telemática y afines	18	No	16	Sí
Ingeniería electrónica, telecomunicaciones y afines	19	No	17	Sí
Ingeniería mecánica y afines	27	Sí	20	Sí
Ingeniería química y afines	56	Sí	52	Sí
Sociología, trabajo social y afines	48	Sí	44	Sí
Ingeniería ambiental (Sin clasificar)	44	Sí	36	Sí

Fuente: Autor

El análisis consolidado de las cohortes históricas entre 2018 y 2023 presentado demuestra un comportamiento altamente heterogéneo en la calidad de los datos, evidenciando que el cumplimiento simultáneo de los criterios del ICFES (**T > 25** y **C > 40%**) es la excepción y no la regla para la institución. Al desglosar el comportamiento por años y Núcleos Básicos del Conocimiento (NBC), se observa lo siguiente:

- **Programas con cumplimiento consistente:** Únicamente los núcleos de **Administración y Derecho y afines** logran cumplir satisfactoriamente con un volumen de evaluados idóneo ($T > 25$) y una tasa de representatividad por cruce superior al umbral mínimo ($C > 40\%$) de manera ininterrumpida a lo largo de toda la serie temporal (2018 a 2023).
- **Programas con fallas críticas recurrentes o estructurales:** El bloque de las ingenierías y las ciencias sociales muestra una vulnerabilidad crítica ante la pérdida de trazabilidad. En la cohorte más reciente (2023), núcleos masivos con alta población observable como **Arquitectura** (201 alumnos), **Ingeniería Civil y afines** (63 alumnos) o **Contaduría Pública** (31 alumnos) fallan rotundamente en el criterio de representatividad ($C > 40\%$), registrando tasas de emparejamiento con Saber 11 inferiores al 5% debido al rezago en las llaves de cruce del ICFES. En el extremo histórico (2018), la degradación es generalizada: con una tasa de éxito global de apenas el 9%, absolutamente todos los NBCs, excepto Administración, incumplen el supuesto de representatividad, dejando con 0 cruces válidos a programas con más de 100 evaluados como Derecho y Arquitectura.
- **Casos por bajo volumen poblacional:** Áreas como **Ingeniería de Sistemas**, **Ingeniería Electrónica** e **Ingeniería Química** registran de forma intermitente periodos sin datos observables (valores en 0) o volúmenes insuficientes ($T > 25$) en años como 2018, 2019, 2020, 2021 y 2022, lo que rompe el supuesto de agregación institucional.

Documentación de conclusiones y recomendaciones

En base al estudio técnico, se sacaron las siguientes conclusiones:

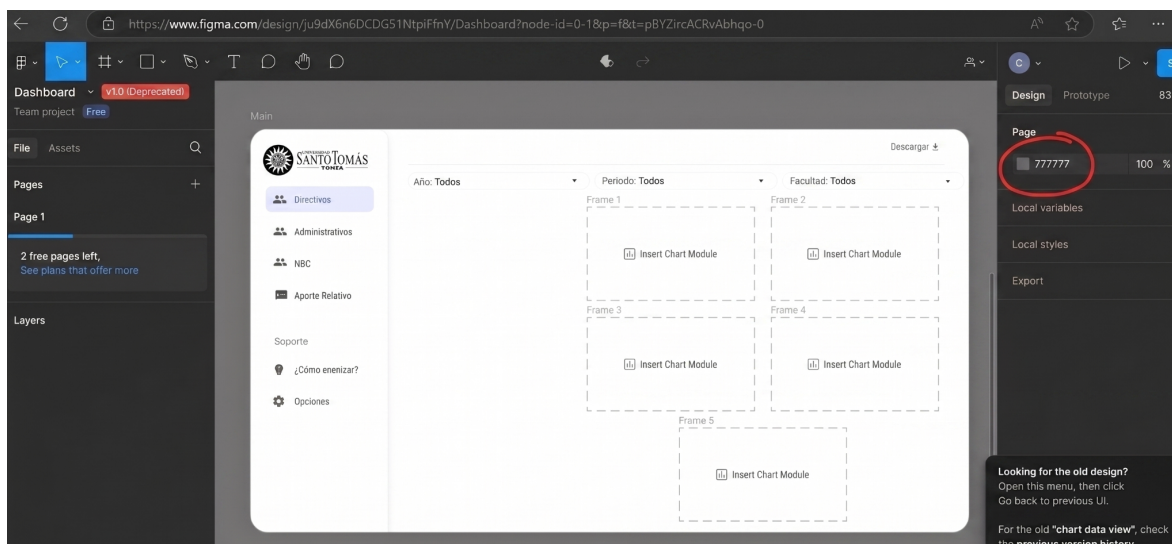
- **Inviabilidad Generalizada Reciente:** La aplicación de modelos predictivos y de valor agregado no es viable técnica ni metodológicamente de manera generalizada para la cohorte más reciente (2023) debido a las fallas estructurales críticas en el porcentaje de cruces con la prueba Saber 11 (27%). Esto responde a la desactualización de las 'llaves' que cruzan los datos más recientes Saber Pro, con los datos históricos Saber 11.
- **Excepciones de Éxito Consistente:** Únicamente los núcleos de Administración y Derecho y afines muestran una estabilidad y representatividad de datos suficiente a lo largo de toda la serie histórica analizada (2018-2023) para ser integrados de forma inmediata en un piloto predictivo institucional.
- **Restricción Multianual:** El requerimiento del ICFES exige que los criterios de representatividad se cumplan simultáneamente en los años objeto de comparación. Al haberse caído los datos de 2023 en las ingenierías y arquitectura, queda bloqueada la posibilidad de emitir informes bienales válidos (ej. periodos combinados 2022-2023) para estas áreas académicas. Restricción que puede disminuir al momento que la entidad ICFES publique las 'llaves de cruce' actualizadas.
- **Acción y Mitigación Urgente:** Se concluye que el problema fundamental no radica en el la cantidad de estudiantes que presentan sus pruebas Saber Pro (pues su segregación por NBC es numeroso y estable), sino en un quiebre en los procesos de recolección, codificación y actualización de los datos históricos debido al rezago que mantiene la entidad ICFES al publicar los datos, y más precisamente las 'llaves de cruce'. Es mandatorio tener presente lo anterior, que no bloquea una futura aplicación del modelo en sí, pero obliga a proceder progresivamente hasta obtener un éxito consistente.

Fase 3. Desarrollo y despliegue

Diseño de Mockup y arquitectura del Dashboard

Diseño Frontend (Figma): Antes de la codificación, se diseñaron mockups (prototipos visuales) utilizando Figma, estructurando la jerarquía del aplicativo web para garantizar una experiencia de usuario (UX) fluida para el personal administrativo no técnico.

Figura 22. Borrador en Figma.



Fuente: Autor

Posteriormente, se creó la arquitectura de los Dashboard, basado en los siguientes filtros:

Filtro de periodo (Año).

Filtro de Departamento.

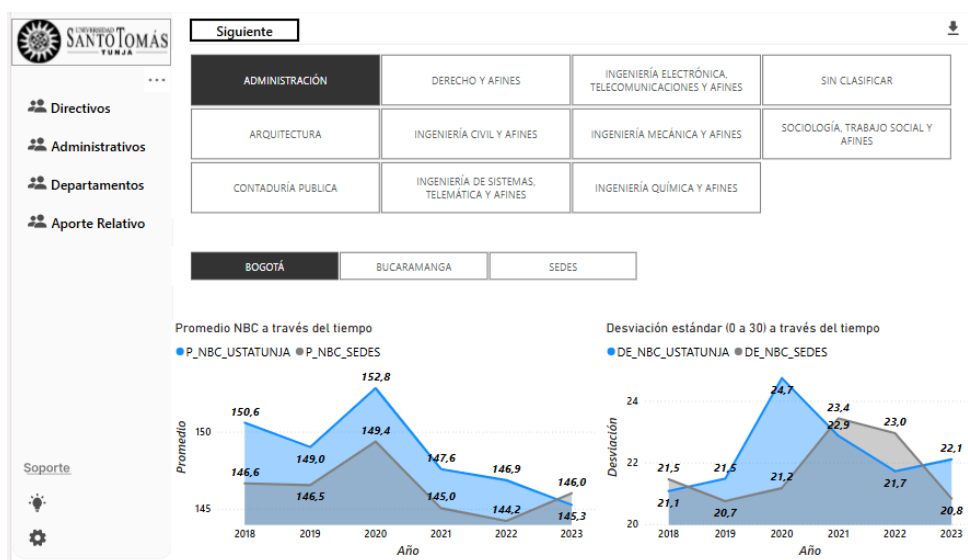
Filtro de Ciudad.

Filtro de NBC.

Filtro comparativo por sede.

Las Figuras que se van a mostrar a continuación son la evidencia en la aplicación de los filtros y las visualizaciones preliminares, que se definieron en base al cumplimiento de los objetivos propuestos.

Figura 23 Sección NBC a través del tiempo (Filtro NBC y Sedes Usta), en producción.



Fuente: Autor

En la Figura 23 se presenta el indicador de Puntaje Global Promedio Institucional y su desviación estándar, segmentado por cohortes históricos. La visualización opera con los filtros activos de Núcleo Básico del Conocimiento (NBC) y Sedes. A diferencia de la vista preliminar (Figura 17), que solo mostraba un nivel de agregación a nivel de Sede y una falta de ubicación, esta versión de producción ya está ubicada e incorpora interactividad con otras vistas, acompañado de más filtros dinámicos.

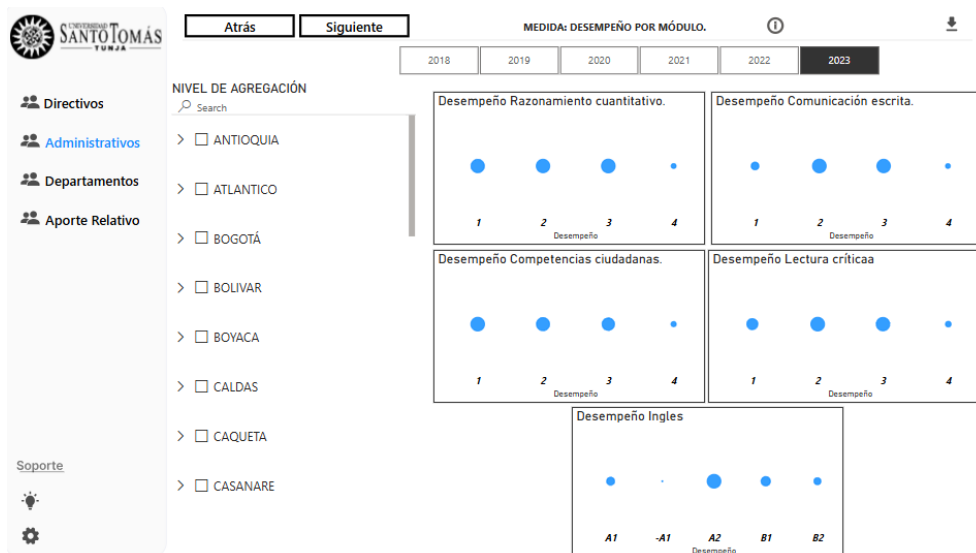
Figura 24 Sección Percentil por NBC (Filtro Sedes Usta), en producción.



Fuente: Autor

La Figura 24 expone el indicador de Distribución del Percentil Institucional frente a las diferentes sedes. Esta vista proporciona a la alta dirección una herramienta clave para procesos de autoevaluación institucional y renovación de registros, midiendo el valor competitivo real de la universidad en el ecosistema de educación superior. A diferencia de la vista preliminar (Figura 18), la Figura 24 muestra la adecuación al Dashboard en producción.

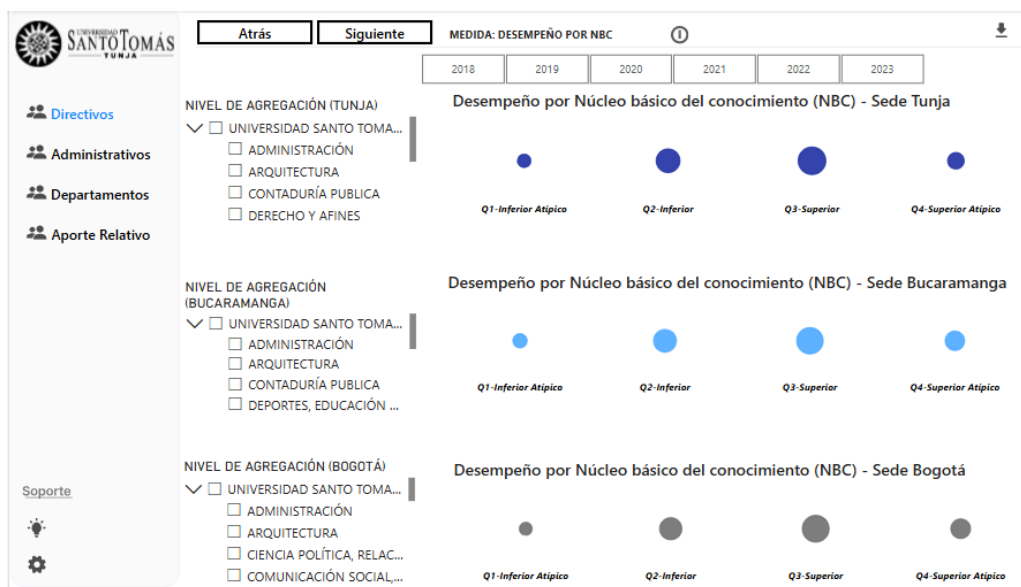
Figura 25 Sección desempeño por competencia (Filtro Nivel de agregación y Año), en producción.



Fuente: Autor

La Figura 25 detalla el indicador de Desempeño por Competencias Genéricas (Lectura Crítica, Razonamiento Cuantitativo, Competencias Ciudadanas, Escritura Proactiva e Inglés). Los filtros activos corresponden a un nivel de agregación histórico a nivel departamental. Respecto a la aproximación preliminar de la Figura 19, la versión final conlleva correcciones ortográficas necesarias y la integración en el mockup de producción.

Figura 26 Sección desempeño por NBC (Filtro Sedes Usta y Año), en producción.



Fuente: Autor

En la Figura 26 se ilustra el indicador de Distribución de Estudiantes por Cuartiles de desempeño. La interfaz mantiene filtros activos para Tipo de Sede y año. La relevancia de esta lectura para la Dirección de Calidad radica en la posibilidad de auditar la equidad formativa, visibilizando si la población estudiantil se concentra en los niveles de desempeño superiores o de rezago. La diferencia fundamental con la fase de diseño inicial (Figura 20) es la integración con las vistas principales, y la adición del filtro histórico de año.

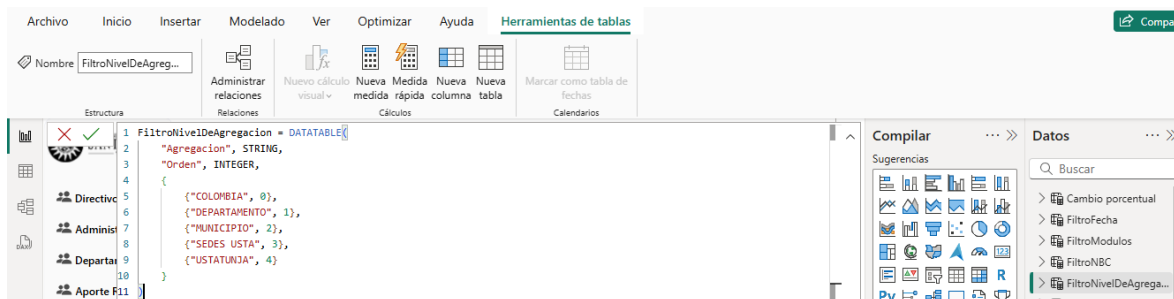
Desarrollo del Dashboard Web

Modelado Relacional: En Power BI, se normaliza el modelo de datos asegurando la correcta cardinalidad e integridad entre las tablas de saber 11, saber pro y consolidado.

Programación DAX: La solución trascendió la simple creación de gráficas. Se requirió la programación de complejas expresiones de análisis de datos (DAX), utilizando funciones de cálculo condicional, evaluación de contextos y transiciones de filtros para que métricas dinámicas (como el Top 5 de puntajes o los percentiles comparativos) respondieran en tiempo real a las interacciones del usuario.

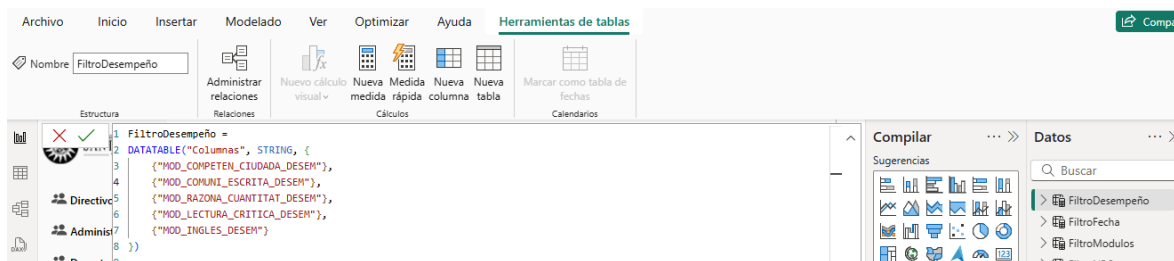
Las siguientes figuras [Figura 27 y Figura 28] nos presentan la creación de algunas de las tablas creadas en el dashboard.

Figura 27. Código Dax para la creación de los niveles de agregación.



Fuente: Autor

Figura 28. Código Dax para la creación del filtro desempeño.



Fuente: Autor

Pruebas funcionales y despliegue

Las pruebas funcionales garantizan la calidad, el rendimiento y la disponibilidad del software entregado.

- Refactorización y Control de Calidad

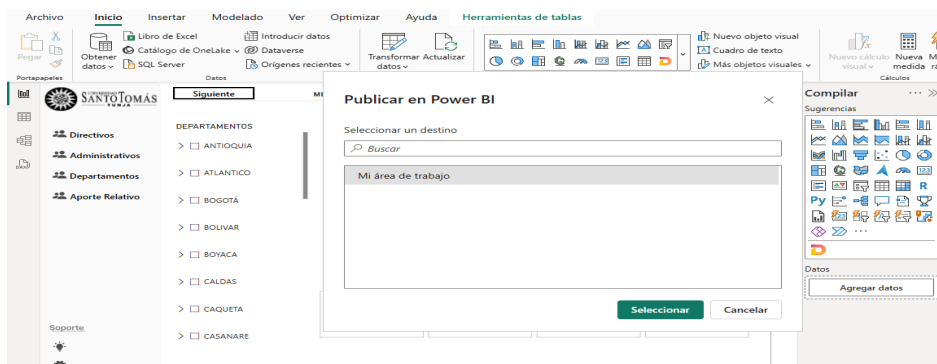
Tal como se documentó en el seguimiento del proyecto, se dedicaron sesiones exhaustivas a la auditoría del código, arreglando redundancias en los scripts de Python y optimizando el modelo en Power BI. Se realizaron pruebas de validación de datos, comparando los resultados arrojados por el Dashboard en Power BI con los PDFs originales emitidos por el ICFES, garantizando una precisión del 100% en la extracción y cálculo de métricas. Adicionalmente, se corrigieron redundancias en el código y se optimizó el tiempo de carga del reporte.

- Integración y Despliegue en la Nube

La arquitectura de solución (Backend/Frontend) se articuló utilizando el ecosistema de Microsoft. Se utilizó SharePoint como repositorio centralizado para los datos, y Power BI Web para el despliegue del aplicativo.

Al culminar las pruebas funcionales, se procedió a desplegar el aplicativo dashboard, y como primer paso, se publicó el proyecto en el Workspace que Power BI Web tiene predeterminado (también es posible crear uno nuevo), que en este caso se llama 'Mi área de trabajo', tal y como se muestra en la siguiente figura.

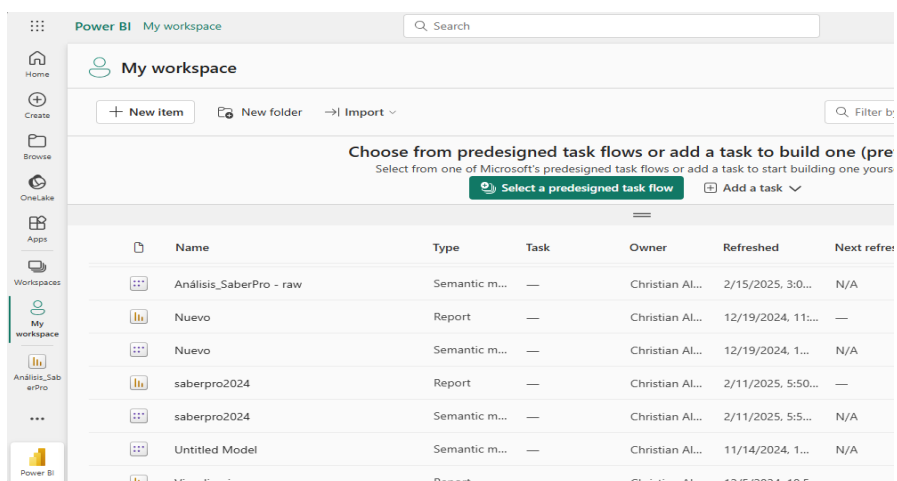
Figura 29. Publicar Dashboard desde el aplicativo local Desktop al aplicativo Web.



Fuente: Autor

Después, nos remitimos a Power BI Web y buscamos el proyecto. [Figura 30]

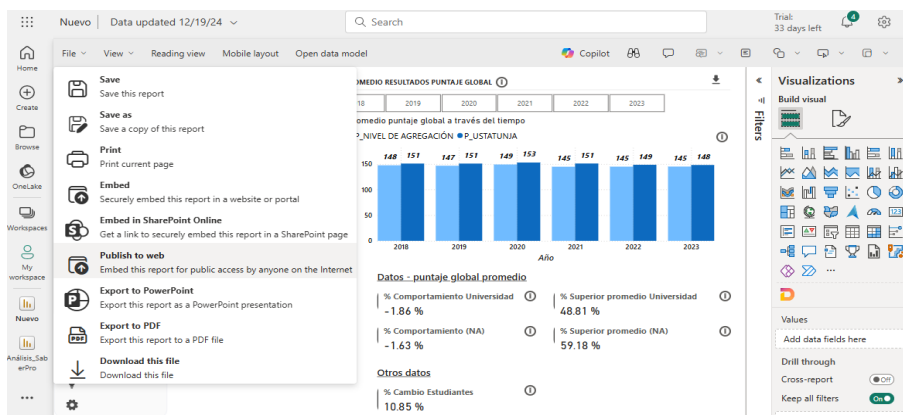
Figura 30. Buscar Dashboard en el aplicativo web.



Fuente: Autor

Ingresamos al proyecto, y nos ubicamos en la opción 'Publish to web' que viene del apartado 'File'. [Figura 31]

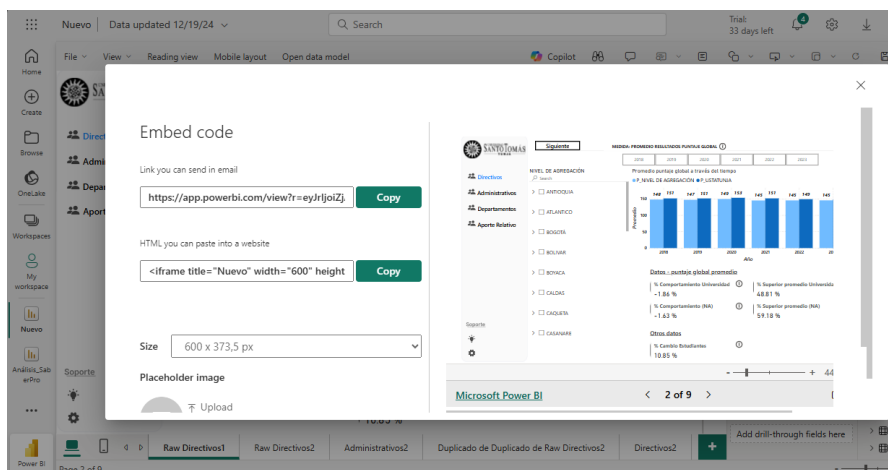
Figura 31. Publicar Dashboard desde el aplicativo Web.



Fuente: Autor

Finalmente, copiamos y guardamos el enlace que nos proporcionan, ya que con ello se comparte el proyecto. [Figura 32]

Figura 32. Muestra del enlace generado cuando el Dashboard se publica en la Web.



Fuente: Autor

Como producto, se finiquitó la puesta a producción de dos Dashboard interactivos, escalables y automatizados en sus particularidades, que centralizan la información académica y reduce significativamente los tiempos de procesamiento manual por parte de la institución. Democratizando el acceso a los datos para la toma de decisiones académicas.

CONCLUSIONES

La culminación de este proyecto de pasantía empresarial ha permitido consolidar una solución analítica institucional integrada y de alta fidelidad para la Dirección de Aseguramiento de la Calidad de la Universidad Santo Tomás. En consistencia con los lineamientos metodológicos propuestos y atendiendo de manera estricta los objetivos específicos del proyecto, se presentan los resultados alcanzados.

Cumplimiento de objetivos

(Objetivo específico 1)

Integración, limpieza y normalización del dataset unificado:

- **Conclusión:** Se logró el diseño y la ejecución exitosa de una infraestructura de datos limpia, consistente y relacional a partir de los microdatos históricos de las pruebas Saber 11 y Saber Pro provistos por el ICFES y los sistemas internos de la Universidad. Mediante técnicas avanzadas de preparación de datos, se mitigaron de raíz las inconsistencias y datos atípicos (como estudiantes sin programa asignado o formatos heterogéneos) y se automatizó el cruce seguro de registros históricos a través del identificador clave (Llaves de cruce).
- **Evidencias:** Scripts estructurados en Python utilizando la librería Pandas para la manipulación matricial de datos, normalización de formatos y estandarización de campos vacíos o nulos.
 - Soporte gráfico detallado en el cuerpo del documento mediante las Figuras 8 a 15, las cuales registran los procesos de uso de expresiones regulares (Regex), verificación de separadores, análisis de PDFs institucionales y resultados finales del cruce relacional.
 - Directorio corporativo centralizado en el SharePoint institucional ("Pasantía 2024"), donde reposa el dataset unificado final.
- **Estado de Cumplimiento:** Cumplido (100%)

(Objetivo específico 2)

Análisis estadístico descriptivo y exploratorio de brechas de rendimiento:

- **Conclusión:** La caracterización estadística exploratoria de las cohortes históricas facilitó el diagnóstico oportuno y la identificación dinámica de brechas de rendimiento académico y patrones de comportamiento por sede y programa institucional. La transición metodológica estática hacia entornos visuales dinámicos dotó a la institución de un puente directo entre el diagnóstico exploratorio y el control minucioso, elevando la profundidad con la que las áreas respectivas analizan competencias genéricas.
- **Evidencias:** Visualizaciones preliminares de exploración de datos configurados para el análisis dinámico de indicadores por año, institución y cohorte académico.
 - Soporte visual en el informe principal a través de las Figuras 17 a 20, que documentan respectivamente las secciones exploratorias de comportamiento de los Núcleos Básicos del Conocimiento (NBC) a través del tiempo, percentiles ponderados por sedes de la USTA y el desempeño relativo por competencia evaluada.
- **Estado de Cumplimiento:** Cumplido (100%)

(Objetivo específico 3)

Evaluación de viabilidad técnica para modelos predictivos y de valor agregado:

- **Conclusión:** Se determinó la viabilidad metodológica para la futura aplicación de analítica avanzada y estimación del valor agregado institucional, concluyendo que el ecosistema cuenta con los supuestos teóricos core (manipulabilidad, no interferencia, supuesto métrico y homogeneidad). No obstante, el estudio arrojó un hallazgo crítico sobre la calidad de la información: existe una fricción operativa en la transición de los documentos de identidad estudiantiles desde Saber 11 hasta Saber Pro, acompañada de una concentración atípica de puntajes cero focalizada en el módulo de Comunicación Escrita (con especial relevancia en el año 2022 con 61 estudiantes afectados), lo que establece la obligatoriedad de auditar y depurar las cohortes antes de implementar modelos predictivos de Machine Learning.

- **Evidencias:** Entregable e informe técnico independiente titulado "Estudio de Viabilidad Técnica: Aplicación de Modelos Predictivos y de Valor Agregado", el cual recopila los análisis de las cohortes 2018-2023 bajo los umbrales de inclusión del ICFES ($T > 25$ estudiantes por NBC y tasa de cruce exitoso ($C > 40\%$) y las conclusiones respectivas.
 - Figura 21 del informe principal, que expone de forma explícita la Matriz de evaluación estratificada por Núcleo Básico del Conocimiento (NBC).
- **Estado de Cumplimiento:** Cumplido (100%)

(Objetivo específico 4)

Diseño, desarrollo e implementación del Dashboard interactivo en producción:

- **Conclusión:** Se culminó con éxito el despliegue de dos Dashboard interactivos, diseñados bajo un enfoque de experiencia de usuario y optimizado para la toma de decisiones estratégicas. El dashboard automatiza la consulta de indicadores de calidad académica complejos (como el cálculo dinámico de percentiles por NBC y sedes a través del lenguaje DAX), descentralizando la información para dotar a las decanaturas y direcciones de una herramienta ágil frente a los retos académicos que la Universidad busca superar.
- **Evidencias:** Dos Dashboard completamente operativos, publicados y desplegados en el entorno en la nube de Power BI Web, accesible de manera interactiva por los usuarios autorizados de la institución.
 - Registro del flujo de diseño de la interfaz mediante la Figura 22 (Borrador de maqueta/Mockup en Figma).
 - Vistas analíticas finales en entorno de producción plenamente implementadas y detalladas en las Figuras 23 a 26, complementadas por las tablas técnicas de navegación, niveles de agregación y filtros de desempeño (Figuras 27 y 28).
 - Manual de Usuario del Dashboard Interactivo, proporcionado como un documento independiente en formato PDF adjunto para guiar la navegación y adopción técnica de la herramienta por parte de los directivos finales.
- **Estado de Cumplimiento:** Cumplido (100%)

Retos y dificultades superadas

Durante la ejecución, el principal reto enfrentado estuvo relacionado con el estado inicial de la información. Se evidenció una carencia importante de estructura y unicidad en los datos históricos, así como el evidente rezago de información, que llevó a un proceso gradual e inevitable para obtener los datos recientes entregados por el ICFES (Resultados individuales en formato .PDF) y proporcionados por la Universidad. Sumado a esto, los dos tipos de fuentes (Históricas y recientes) requirieron procesos ETL particulares. Enfrentar esta limitación técnica y de tiempo obligó a replantear las prioridades, demostrando que el desarrollo de software debe alinearse dinámicamente con las necesidades organizacionales para ofrecer soluciones inmediatas.

Tabla 5. Retos técnicos y metodológicos..

Retos técnicos / Metodológicos	Acción Adoptada (Mitigación)	Resultado Concreto Alcanzado
Rezago de la entidad en la entrega de sus bases de datos.	Se tabularon los datos recuperados y se documentó formalmente que la mitigación de este rezago es temporal y supeditada a los tiempos de la entidad ICFES.	Consolidación y blindaje de la serie histórica viable (2018-2023) para futuros estudios.
Necesidad de extraer archivos PDF de resultados individuales mediante una solución que no sea dependiente de conocimientos técnicos avanzados ni genere	Diseño de un pipeline de procesamiento en Python basado netamente en expresiones regulares (Regex), el cual captura la información de forma directa aprovechando	Extracción automatizada del 100% de la información e ingesta limpia en el SharePoint institucional (" <i>Pasantía 2024</i> "), garantizando

dependencias a librerías externas complejas.	patrones sintácticos fijos en el texto nativo del PDF.	una solución libre de dependencias complejas.
Baja representatividad muestral por NBC en cohortes específicas.	Reorientación estratégica del proyecto hacia el fortalecimiento de la analítica descriptiva reciente e histórica y el despliegue del Dashboard correspondiente.	Despliegue del dashboard en Power BI Web, dotando al área correspondiente de una herramienta operativa para la toma de decisiones y planes de mejoramiento pedagógico.

Fuente: Autor

Aprendizaje y competencias desarrolladas

La pasantía empresarial permitió consolidar las competencias profesionales mediante un contraste directo entre los fundamentos teóricos del programa académico y las exigencias del entorno laboral real. Esta articulación se resume a continuación:

- **Componentes académicos aplicados:** Los conocimientos de las asignaturas de Bases de Datos (para el modelado de datos y reinterpretación del lenguaje SQL en un lenguaje DAX), Ingeniería de Software (para el diseño arquitectónico de la solución mediante Scripts python en ambientes locales), Machine Learning (Para la correcta implementación de procesos ETL enfocados en el análisis descriptivo y exploratorio de grandes cantidades de datos) y Estadística sirvieron como cimiento conceptual y metodológico indispensable para estructurar el proyecto.
- **Competencias adquiridas exclusivamente en la pasantía:** La práctica real exigió afrontar retos que superaron el alcance tradicional de las aulas. Entre ellos, la construcción de flujos ETL reales sobre información heterogénea, la extracción automatizada mediante expresiones regulares (Regex) sobre archivos PDF individuales, el dominio de funciones analíticas complejas (DAX) en Power BI, el

entendimiento de la normativa institucional del ICFES y el manejo técnico de bases de datos incompletas o anonimizadas.

A nivel estratégico, la experiencia afianzó competencias blandas críticas como la adaptabilidad ante escenarios técnicos flexibles, el criterio profesional para evaluar la calidad de la información y la capacidad de traducir problemas operativos en soluciones tecnológicas funcionales de impacto inmediato.

Impacto, automatización y beneficios concretos

El impacto de la intervención en la Universidad fue significativo y tangible. La implementación del proceso ETL (Extracción, Transformación y Carga) junto con la automatización de la lectura de resultados (PDFs) redujo drásticamente la carga operativa institucional. Esto mitigó los errores humanos asociados al procesamiento manual y logró consolidar la información de los años 2023 y 2024 en un repositorio estructurado. Como resultado, la solución democratizó el acceso a los datos, beneficiando directamente a los directivos y tomadores de decisiones de los programas académicos, quienes ahora pueden contar con un sistema intuitivo para evaluar el desempeño de los estudiantes frente a métricas nacionales (como el percentil por NBC).

Proyección institucional

Finalmente, la culminación de la fase de 'Evaluación de viabilidad técnica' dejó un impacto a largo plazo. Al organizar los datos y diagnosticar el estado actual de la información, se estableció el prerequisite fundamental que garantiza la calidad técnica para que la institución pueda ejecutar procesos de analítica predictiva y Machine Learning en el corto y mediano plazo.

RECOMENDACIONES Y TRABAJOS FUTUROS

Teniendo en cuenta que la primera versión de la solución fue estabilizada y desplegada hacia octubre de 2024, se sugieren las siguientes líneas de trabajo para la mejora continua del aplicativo:

- **Migración de Repositorio a Motor Relacional Específico:** Se recomienda migrar el almacenamiento de los archivos procesados (actualmente apoyado en ecosistemas locales/SharePoint) hacia un gestor de bases de datos relacional robusto (como PostgreSQL). Esto dotará al aplicativo web de mayor velocidad de respuesta y seguridad a medida que aumente la volumetría de datos.
- **Automatización Total del Flujo de Ingesta:** Incorporar un flujo integrado (por ejemplo, mediante Power Automate) que permita que la plataforma detecte automáticamente la llegada de nuevos PDFs de resultados desde el Icfes, ejecute el script de Python en la nube y actualice el tablero de Power BI sin necesidad de intervención manual por parte de un administrador.
- **Refactorización en busca de innovación:** Utilizar LLMs para la extracción de los datos como futuro refactor, alineándose así a las nuevas innovaciones.
- **Desarrollo del Modelo Predictivo:** Aprovechando que la información ahora se encuentra limpia, estandarizada y normalizada, la Universidad puede retomar el objetivo original del proyecto, utilizando este dataset como base de estudio y entrenamiento para un modelo que logre identificar patrones de desempeño y proyectar resultados futuros con alta precisión.

ANEXOS

Los anexos correspondientes al desarrollo de este proyecto han sido debidamente proporcionados como documentos independientes y adjuntos al presente informe. A continuación, se presenta la relación de los entregables:

- Anexo A. Código de limpieza e integración para datos históricos. (Archivo .IPYNB adjunto).
- Anexo B. Código de limpieza e integración para datos recientes que vienen en formato PDF (Archivo .IPYNB adjunto).
- Anexo C. Aplicativo Dashboard de datos históricos. (Archivo .PBIX adjunto).
- Anexo D. Aplicativo Dashboard de datos recientes. (Archivo .PBIX adjunto).
- Anexo E. Manual de Usuario Aplicativo Dashboard de datos históricos (2018-2023) (Archivo PDF adjunto).
- Anexo F. Manual de Usuario Aplicativo Dashboard de datos recientes (Archivo PDF adjunto).
- Anexo G. Documentación de estudio de viabilidad técnica para la implementación de un modelo predictivo sobre los resultados Saber Pro. (Archivo PDF adjunto).
- Anexo H. Ubicación del Repositorio de Datos y Códigos: SharePoint Institucional / Directorio Pasantía 2024.

REFERENCIAS

- Departamento Administrativo Nacional de Estadística (DANE). (s. f.). Archivo Nacional de Datos (ANDA). Recuperado el 26 de junio de 2026, de <https://www.dane.gov.co/>
- IBM. (s. f.). What Is Business Intelligence (BI)? Recuperado el 26 de junio de 2026, de <https://www.ibm.com/think/topics/business-intelligence>
- Instituto Colombiano para la Evaluación de la Educación (ICFES). (s. f.-a). Acerca del examen Saber 11°. Recuperado el 26 de junio de 2026, de <https://www.icfes.gov.co/acerca-del-examen-saber-11>
- Instituto Colombiano para la Evaluación de la Educación (ICFES). (s. f.-b). Acerca del examen Saber Pro. Recuperado el 26 de junio de 2026, de <https://www.icfes.gov.co/acerca-del-examen-saber-pro>
- Instituto Colombiano para la Evaluación de la Educación (ICFES). (2021). Guía de orientación para la interpretación de resultados del examen Saber Pro. ICFES. Recuperado el 26 de junio de 2026, de <https://www.icfes.gov.co/>
- MDN Web Docs. (2024). Regular expressions. Recuperado el 26 de junio de 2026, de https://developer.mozilla.org/en-US/docs/Web/JavaScript/Guide/Regular_expressions
- Microsoft Learn. (s. f.). Data Analysis Expressions (DAX) Reference. Recuperado el 26 de junio de 2026, de <https://learn.microsoft.com/es-es/dax/>
- Ministerio de Educación Nacional de Colombia. (s. f.-a). Exámenes de Estado. Recuperado el 26 de junio de 2026, de <https://www.mineduccion.gov.co/>
- Ministerio de Educación Nacional de Colombia. (s. f.-b). Glosario SNIES. Recuperado el 26 de junio de 2026, de <https://snies.mineduccion.gov.co/portal/DOCUMENTOS/Glosario/>