

Proyecto de Integración y Automatización de Datos para IA para el proyecto Brújula Uni: Facilitando Decisiones Informadas en la Educación Superior.

Alexander Victoria¹, Elthon Brayan Marquez², Maria Alejandra Fandiño³

¹Facultad de Ingeniería y Ciencias Básicas

Universidad Central

Maestría en Analítica de Datos

Automatización e Integración de Datos para IA

Bogotá, Colombia

{¹avictoriag,²emarquezb}@correo1.com, ³mfandinom@correo2.com

November 25, 2023

Contents

1	Introducción (Max 250 Palabras) - (<i>Primera entrega</i>)	3
2	Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA (Max 500 Palabras) - (<i>Primera entrega</i>)	4
2.1	Titulo del proyecto de investigación (Max 100 Palabras) - (<i>Primera entrega</i>)	5
2.2	Objetivo general (Max 100 Palabras) - (<i>Primera entrega</i>)	5
2.2.1	Objetivos especificos (Max 100 Palabras) - (<i>Primera entrega</i>)	5
2.3	Alcance (Max 200 Palabras) - (<i>Primera entrega</i>)	5
2.4	Pregunta de investigación (Max 100 Palabras) - (<i>Primera entrega</i>)	6
2.5	Hipotesis (Max 100 Palabras) - (<i>Primera entrega</i>)	6
3	Reflexiones sobre el origen de datos e información (Max 400 Palabras) - (<i>Primera entrega</i>)	6
3.1	¿Cual es el origen de los datos e información ? (Max 100 Palabras) - (<i>Primera entrega</i>)	7
3.2	¿Cuales son las consideraciones legales o éticas del uso de la información? (Max 100 Palabras) - (<i>Primera entrega</i>)	7

3.3	¿Cuales son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA? (Max 100 Palabras) - (<i>Primera entrega</i>)	7
3.4	¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto? (Max 100 Palabras) - (<i>Primera entrega</i>)	8
4	Diseño de integración y Automatización de Datos para IA (Diagrama) (<i>Primera entrega</i>)	8
5	integración de Datos (<i>Segunda entrega</i>)	9
6	Automatización de Datos (<i>Segunda entrega</i>)	13
7	IA (<i>Segunda entrega</i>)	19
8	Proximos pasos (<i>Tercera entrega</i>)	20
9	Lecciones aprendidas (<i>Tercera entrega</i>)	21
10	Bibliografía	22

1 Introducción (Max 250 Palabras) - (*Primera entrega*)

En la actualidad, el acceso a la educación superior es una decisión crítica que puede determinar el futuro de los estudiantes. Sin embargo, la elección de una universidad y una carrera se ha vuelto compleja y abrumadora debido a la diversidad de opciones y la abundancia de información disponible sobre carreras, universidades, etc. Los estudiantes se enfrentan a una marea de información, desde resultados académicos hasta costos de matrícula y opiniones de otros estudiantes. Para realizar una elección vocacional acertada, se requiere de un determinado nivel de desarrollo de madurez vocacional que permita establecer objetivos profesionales que sean plenos para la persona y para la demanda del mercado laboral (López Fernández Sánchez Herrera, 2018).

Surge así la necesidad de crear una herramienta que simplifique este proceso y permita a los futuros estudiantes tomar decisiones educativas más informadas.

La elección de una carrera universitaria es una de las decisiones más significativas en la vida de un estudiante que quiera continuar con la educación superior, implicando una inversión considerable de tiempo y recursos, así como la construcción de un camino hacia metas personales y profesionales. En la mayoría de los casos, los jóvenes a la hora de definir qué estudiar toman en cuenta factores como el mercado laboral que encontrarán al graduarse de la universidad, el salario de enganche” (Cardenas, 2015) La pregunta clave que enfrentan los estudiantes es: ¿Cómo puedo tomar una decisión tan import- ante de manera informada y precisa?

Para abordar esta pregunta, se propone el desarrollo de un proyecto de automatización centrado en la creación de un comparador de universidades por carrera. Esta plataforma busca proporcionar a los futuros estudiantes una herramienta integral y fácil de usar que les permita evaluar de manera efectiva las opciones disponibles en función de múltiples factores clave. Desde el desempeño académico de las universidades hasta los detalles específicos de los programas de estudio y las experiencias de otros estudiantes, esta herramienta busca ofrecer una visión completa y precisa de las alternativas educativas que actualmente dispone nuestro país.

En un mundo impulsado por la información y la tecnología, es esencial contar con herramientas que simplifiquen la toma de decisiones y ayuden a los estudiantes a plantear una perspectiva clara de la educación superior. El presente proyecto busca abordar este desafío, brindando a los futuros estudiantes una brújula confiable para explorar sus opciones académicas y tomar decisiones que estén en consonancia con sus objetivos y aspiraciones, Este apoyo en la orientación profesional constituye “una actividad de información y asesoramiento que ayuda al estudiante a realizar una decisión vocacional coherente, una buena elección profesional” (Tintaya Condori, 2016).

A largo plazo, se espera que esta herramienta pueda llegar a impactar en los indicadores de deserción en la educación superior que tal como reportó el ministerio de educación nacional para el 2021 se encontraba el porcentaje de deserción anual en un 10,08 (Spadies - Estadísticas de deserción, 2023)

2 Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA (Max 500 Palabras) - (Primera entrega)

El proyecto de investigación "Brújula Uni: Facilitando Decisiones Informadas en la Educación Superior" se enfoca en el desarrollo de un comparador de universidades por carrera. Su objetivo principal es proporcionar a los estudiantes una herramienta robusta y de fácil acceso que les permita evaluar de manera efectiva las múltiples opciones disponibles en función de una amplia variedad de factores clave, que van desde el desempeño académico de las universidades hasta detalles específicos de los programas de estudio y las experiencias de otros estudiantes.

Para lograr este propósito, se plantean los siguientes objetivos que guiarán el desarrollo y la implementación de la plataforma "Brújula Uni". En primer lugar, se llevará a cabo la obtención y almacenamiento exhaustivo de datos actualizados sobre las carreras de educación, incluyendo resultados ponderados de las pruebas Saber Pro hasta el año 2022, costos de pregrado, duración de semestre, modalidad de estudio, entre otros. Este proceso se realizará a través de fuentes de datos abiertas y de acceso público, garantizando la legalidad y ética en el uso de la información, adicionalmente se implementará un proceso de web scraping para obtener información no estructurada sobre los planes de estudio de las carreras y las opiniones de estudiantes. Esta técnica permitirá enriquecer la información disponible y proporcionar una visión más completa de cada programa académico.

En segundo lugar, se diseñará y aplicará un sistema de ponderación a través de algoritmos de clasificación, que evaluará y asignará valores a las diversas características de las carreras y las universidades. Esto permitirá a los futuros estudiantes tener una visión equilibrada y objetiva de las opciones disponibles, facilitando así la comparación y toma de decisiones.

En tercer lugar, se creará una interfaz orientada al usuario, intuitiva y de fácil navegación. Esta plataforma será el punto de acceso principal, donde podrán seleccionar una carrera y comparar universidades en función de las características ponderadas.

Es importante destacar que este proyecto se centra en simplificar la toma de decisiones en el ámbito de la educación superior y no pretende reemplazar la orientación profesional o académica proporcionada por expertos en el campo.

2.1 Título del proyecto de investigación (Max 100 Palabras) - (*Primera entrega*)

Brújula Uni: Facilitando Decisiones Informadas en la Educación Superior.

2.2 Objetivo general (Max 100 Palabras) - (*Primera entrega*)

Crear un sistema con procesos automatizados para evaluar universidades de ciencias de la educación en Bogotá, teniendo en cuenta resultados académicos, características de programas, puntajes y opiniones estudiantiles.

2.2.1 Objetivos específicos (Max 100 Palabras) - (*Primera entrega*)

- Obtener y almacenar de manera exhaustiva datos actualizados sobre las carreras de educación para la ciudad de bogotá, incluyendo resultados ponderados de las pruebas Saber Pro hasta 2022, costos de pregrado, duración de semestre, modalidad de estudio, plan de estudio, entre otras.
- Diseñar y aplicar un sistema de ponderación que evalúe y asigne valores a las diversas características de las carreras y las universidades, teniendo en cuenta su relevancia en la toma de decisiones de los estudiantes.
- Crear un esquema para desarrollar una interfaz orientada al usuario que pueda ayudar a los futuros estudiantes seleccionar una carrera y comparar universidades en función de las características ponderadas.
- Implementar un proceso de web scraping para obtener información no estructurada sobre los pensums académicos de las carreras, enriqueciendo de esta manera la información disponible y brindando una visión más completa de cada programa académico.

2.3 Alcance (Max 200 Palabras) - (*Primera entrega*)

Este proyecto busca desarrollar un sistema que centralice información crucial sobre carreras y universidades, facilitando así la toma de decisiones en la elección de programas académicos para la educación superior.

Para lograr esta misión, se llevará a cabo la recopilación y el almacenamiento de datos detallados sobre universidades que dictan las carreras de licenciatura en la ciudad de Bogotá, dentro de esta información, en una primera etapa, se tendrá en cuenta los resultados de las pruebas Saber Pro entre 2018 y 2022, el número de inscripciones a primer semestre en el año 2022 y los aspectos relacionados con los precios de matrícula.

Se empleará un sistema de ponderación para clasificar y evaluar estos factores. Además, se diseñará la interfaz de usuario para facilitar la comparación de opciones.

Dentro de este proceso, se empleará la técnica de web scraping para enriquecer aún más la información disponible con detalles sobre planes de estudio.

2.4 Pregunta de investigación (Max 100 Palabras) - (*Primera entrega*)

¿En qué medida puede la automatización de datos y la integración de información sobre carreras y universidades, mejorar la calidad de la información disponible para los estudiantes que buscan orientación en la elección de su educación superior?

2.5 Hipotesis (Max 100 Palabras) - (*Primera entrega*)

Con este proyecto, se espera demostrar que una manera condensada de presentar la información de las diferentes puntuaciones que obtienen las universidades según las variables a evaluar por parte del futuro estudiante, permitirá realizar un análisis más detallado lo cual le dará al usuario un marco conceptual más amplio para tomar decisiones más informadas

3 Reflexiones sobre el origen de datos e información (Max 400 Palabras) - (*Primera entrega*)

Para llevar a cabo este proyecto, se recurrió a diversas fuentes de información, todas ellas de acceso público, ya que fueron proporcionadas por distintas entidades gubernamentales. Inicialmente, se empleó la base de datos con los resultados

anonimizados de la prueba Saber Pro del período comprendido entre los años 2018 a 2022. Posteriormente, se utilizó otra base de datos que contiene a

los estudiantes matriculados en educación superior para Colombia en 2022; esta base es suministrada por el Ministerio de Educación Nacional del país basada en los reportes de las propias instituciones de educación superior a través del Sistema Nacional de Información de la Educación Superior (SNIES). Finalmente, se incorporó la base de datos del SNIES que proporciona información

sobre los programas académicos ofrecidos en el área de ciencias de la educación en el país. Sobre todas las bases de datos, se realizó un filtro a la información

seleccionando únicamente los programas profesionales de ciencias de educación. Es importante destacar que durante el proceso de análisis, se identificaron

datos faltantes en algunos registros, particularmente en los resultados de las pruebas Saber Pro y en la información de los costos de matrícula de algunas universidades.

3.1 ¿Cual es el origen de los datos e información ? (Max 100 Palabras) - (*Primera entrega*)

Las bases de datos fueron recopiladas de las siguientes fuentes de información:

- **Datos abiertos:** Según el Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC), la información pública se dispone en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. (MinTIC, [2021])
- **SNIES:** Según el Ministerio de educación nacional, el Sistema Nacional de Información de la Educación Superior (SNIES), es un sistema que ha sido creado para responder a las necesidades de información de la educación superior en Colombia. En este sistema se recopila y organiza la información relevante sobre la educación superior que permite hacer planeación, monitoreo, evaluación, asesoría, inspección y vigilancia del sector. (Ministerio de Educación Nacional, [2021])

3.2 ¿Cuales son las consideraciones legales o éticas del uso de la información? (Max 100 Palabras) - (*Primera entrega*)

Los datos abiertos son información pública que se puede reutilizar bajo licencia abierta y sin restricciones legales en Colombia. Esta regulación está respaldada por la Ley 1712 de 2014, la cual establece la obligación de las entidades públicas de poner a disposición datos y define los datos abiertos en su numeral sexto como: "Todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos". (Ley 1712 de 2014, Colombia)

3.3 ¿Cuales son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA? (Max 100 Palabras) - (*Primera entrega*)

Los retos en la Integración y Automatización de Datos para el desarrollo del proyecto incluyen la complejidad y calidad de la información, la gestión de procesos de web scraping, la integración de múltiples fuentes y el diseño preciso de ponderaciones. Estos desafíos deben abordarse para garantizar que "Brújula Uni" ofrezca a los futuros estudiantes información precisa y relevante.

3.4 ¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto? (Max 100 Palabras) - (Primera entrega)

Esperamos que la automatización y la integración de datos mejore la calidad de la información para los estudiantes que buscan orientación en la elección de su universidad, esto simplificará la recopilación y actualización de datos, ahorrando tiempo y recursos en comparación con consultas sobre carreras efectuadas de forma manual. Además, el uso de técnicas de web scraping enriquecerá la información disponible, proporcionando detalles sobre planes de estudio para ofrecer una visión más completa de cada programa académico. En última instancia, esta implementación sentará las bases técnicas y la arquitectura para el desarrollo de la plataforma "Brújula Uni".

4 Diseño de integración y Automatización de Datos para IA (Diagrama) (Primera entrega)

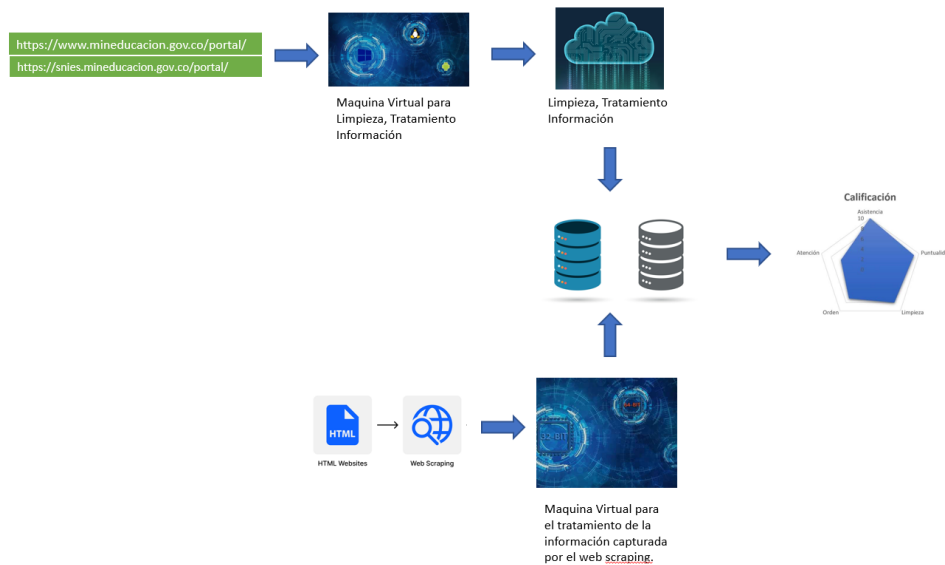


Figure 1: Diseño de integración.

5 integración de Datos *(Segunda entrega)*

Para la integración de datos de este proyecto se utilizaron dos sistemas de administración de bases de datos.

Para los datos relacionales se trabajó con MySQL donde se cargaron las bases con los resultados de la prueba saber pro, con los datos de los programas académicos y con los datos de matriculados bajo el siguiente modelo lógico -relacional.

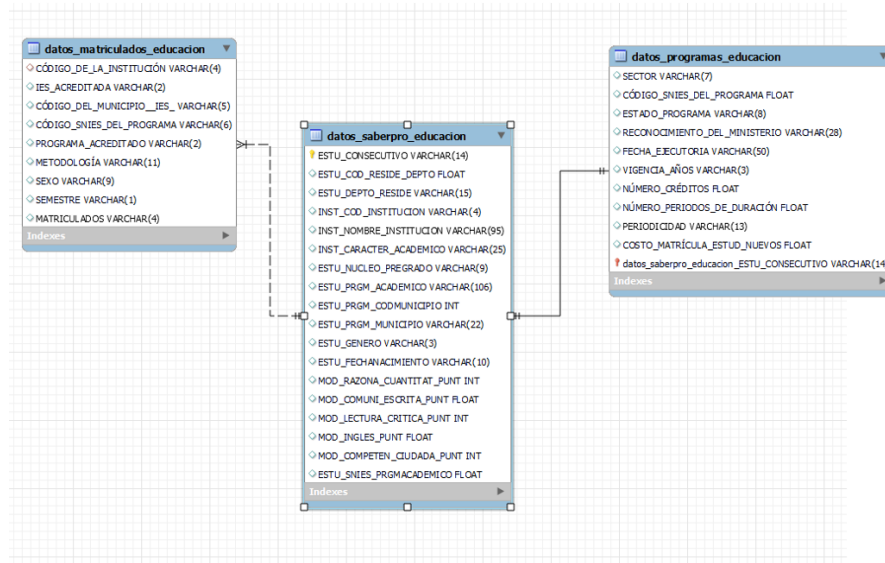


Figure 2: Modelo lógico

A continuación, se anexan las imágenes de las tablas de las bases de datos en Mysql

1 • SELECT * FROM DATOS_SABERPRO_EDUCACION;

ESTU_CONSECUTIVO	ESTU_COD_RESIDE_DEPTO	ESTU_DEPTO_RESIDE	INST_COD_INSTITUCION	INST_NOMBRE_INSTITUCION	INST_CARACTER_ACADEMICO	ESTU_NUCLEO_PREGRADO
EK201830001697	23	CORDOBA	1113	UNIVERSIDAD DE CORDOBA-MONTERIA	UNIVERSIDAD	EDUCACIÓN
EK201830002147	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002157	52	NARIÑO	1206	UNIVERSIDAD DE NARIÑO-PASTO	UNIVERSIDAD	EDUCACIÓN
EK201830002160	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002168	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002174	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002221	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002233	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002243	11	BOGOTÁ	1806	UNIVERSIDAD LIBRE-BOGOTÁ D.C.	UNIVERSIDAD	EDUCACIÓN
EK201830002247	52	NARIÑO	1206	UNIVERSIDAD DE NARIÑO-PASTO	UNIVERSIDAD	EDUCACIÓN
EK201830002253	50	META	1119	UNIVERSIDAD DE LOS LLANOS-VILLAVICEN...	UNIVERSIDAD	EDUCACIÓN
EK201830002261	11	BOGOTÁ	2710	FUNDACION UNIVERSITARIA MONSERRATE...	INSTITUCIÓN UNIVERSITARIA	EDUCACIÓN
EK201830002295	50	META	1119	UNIVERSIDAD DE LOS LLANOS-VILLAVICEN...	UNIVERSIDAD	EDUCACIÓN
EK201830002315	11	BOGOTÁ	2710	FUNDACION UNIVERSITARIA MONSERRATE...	INSTITUCIÓN UNIVERSITARIA	EDUCACIÓN
EK201830002317	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002326	50	META	1119	UNIVERSIDAD DE LOS LLANOS-VILLAVICEN...	UNIVERSIDAD	EDUCACIÓN
EK201830002328	25	CUNDINAMARCA	2710	FUNDACION UNIVERSITARIA MONSERRATE...	INSTITUCIÓN UNIVERSITARIA	EDUCACIÓN
EK201830002335	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002337	76	VALLE	1203	UNIVERSIDAD DEL VALLE-CALI	UNIVERSIDAD	EDUCACIÓN
EK201830002346	50	META	1119	UNIVERSIDAD DE LOS LLANOS-VILLAVICEN...	UNIVERSIDAD	EDUCACIÓN
EK201830002357	18	CAQUETA	1115	UNIVERSIDAD DE LA AMAZONIA-FLORENCIA	UNIVERSIDAD	EDUCACIÓN
EK201830002358	50	META	1119	UNIVERSIDAD DE LOS LLANOS-VILLAVICEN...	UNIVERSIDAD	EDUCACIÓN

Figure 3: Tabla Saberpro

CODIGO PROYECTO BASES... SQL File 3" x

Limit to 1000 rows

1 • SELECT * FROM DATOS_MATRICULADOS_EDUCACION;

CÓDIGO_DE_LA_INSTITUCIÓN	IES_AREDITADA	CÓDIGO_DEL_MUNICIPIO_IES	CÓDIGO_SNIES_DEL_PROGRAMA	PROGRAMA_AREDITADO	METODOLOGÍA	SEXO	SEMESTRE	MATRICULADOS
1101	SI	11001	52746	NO	Presencial	Masculino	1	21
1101	SI	11001	52746	NO	Presencial	Masculino	2	33
1101	SI	11001	52746	NO	Presencial	Femenino	1	30
1101	SI	11001	52746	NO	Presencial	Femenino	2	32
1101	SI	11001	52978	NO	Presencial	Masculino	1	34
1101	SI	11001	52978	NO	Presencial	Masculino	2	29
1101	SI	11001	52978	NO	Presencial	Femenino	1	30
1101	SI	11001	52978	NO	Presencial	Femenino	2	30
1101	SI	11001	53264	NO	Presencial	Masculino	1	2
1101	SI	11001	53264	NO	Presencial	Masculino	2	6
1101	SI	11001	53264	NO	Presencial	Femenino	1	4
1101	SI	11001	53264	NO	Presencial	Femenino	2	5
1101	SI	11001	55136	NO	Presencial	Masculino	1	22
1101	SI	11001	55136	NO	Presencial	Masculino	2	23
1101	SI	11001	55136	NO	Presencial	Femenino	1	19
1101	SI	11001	55136	NO	Presencial	Femenino	2	17
1101	SI	11001	101795	NO	Presencial	Masculino	1	9
1101	SI	11001	101795	NO	Presencial	Masculino	2	10
1101	SI	11001	101795	NO	Presencial	Femenino	1	9
1101	SI	11001	101795	NO	Presencial	Femenino	2	8
1101	SI	11001	106672	NO	Presencial	Masculino	1	8
1101	SI	11001	106672	NO	Presencial	Masculino	2	14

Figure 4: Tabla Matriculados

SECTOR	CODIGO_SINIES_DEL_PROGRAMA	ESTADO_PROGRAMA	RECONOCIMIENTO_DEL_MINISTERIO	FECHA_EJECUTORIA	VIGENCIA_AÑOS	NUMERO_CREDITOS	NUMERO_PERIODOS_DE_DURACION	PERIODICIDAD	COSTO_MATRICULA_ESTUD_NUEVOS
Oficial	146	Activo	Acreditación de alta calidad	2023-03-03 00:00:00	6	170	10	Semestral	1284790
Oficial	147	Activo	Acreditación de alta calidad	2023-03-03 00:00:00	6	157	10	Semestral	1284790
Oficial	148	Inactivo	0	0	0	0	8	Semestral	0
Oficial	149	Inactivo	0	0	0	160	10	Semestral	0
Oficial	150	Inactivo	Acreditación de alta calidad	0	6	160	10	Semestral	0
Oficial	151	Activo	Acreditación de alta calidad	2023-03-03 00:00:00	6	168	10	Semestral	1284790
Oficial	152	Inactivo	0	0	0	0	10	Semestral	0
Oficial	153	Inactivo	Acreditación de alta calidad	0	6	160	10	Semestral	0
Oficial	154	Inactivo	Acreditación de alta calidad	0	6	160	10	Semestral	0
Oficial	155	Activo	Acreditación de alta calidad	2023-05-19 00:00:00	4	160	10	Semestral	1284790
Oficial	156	Activo	Acreditación de alta calidad	2022-05-24 00:00:00	6	160	10	Semestral	1284790
Oficial	157	Activo	Acreditación de alta calidad	2020-11-18 00:00:00	4	160	10	Semestral	1284790
Oficial	158	Activo	Acreditación de alta calidad	2018-04-25 00:00:00	6	160	10	Semestral	1284790
Oficial	159	Activo	Acreditación de alta calidad	2019-06-12 00:00:00	4	134	8	Semestral	1284790
Oficial	161	Activo	Registro calificado	2016-10-10 00:00:00	7	24	2	Semestral	0
Oficial	162	Activo	Registro calificado	2016-10-10 00:00:00	7	23	2	Semestral	5000000
Oficial	163	Inactivo	Registro calificado	2006-04-28 00:00:00	7	36	3	Semestral	0
Oficial	165	Inactivo	0	2016-10-10 00:00:00	0	24	2	Semestral	0
Oficial	166	Inactivo	0	0	0	0	4	Semestral	0
Oficial	167	Inactivo	0	0	0	0	4	Semestral	0
Oficial	169	Activo	Acreditación de alta calidad	2023-04-25 00:00:00	6	48	4	Semestral	6000000
Oficial	170	Activo	Acreditación de alta calidad	2023-03-03 00:00:00	6	50	4	Semestral	5000000
Oficial	171	Inactivo	0	0	0	0	4	Semestral	0
Oficial	177	Inactivo	Acreditación de alta calidad	2017-05-11 00:00:00	6	45	4	Semestral	5000000

Figure 5: Tabla programas

Para gestionar datos no relacionales, se optó por emplear MongoDB como sistema de gestión de bases de datos. En este entorno, se procedió a la carga de los documentos extraídos previamente mediante la técnica de web scraping. Este proceso se llevó a cabo siguiendo un modelo lógico diseñado específicamente para la tarea en cuestión.

MongoDB, una base de datos NoSQL ampliamente reconocida, se eligió por su capacidad para manejar datos semi-estructurados o no relacionales de manera eficiente. A diferencia de los sistemas de gestión de bases de datos relacionales tradicionales, MongoDB almacena los datos en forma de documentos JSON (JavaScript Object Notation), lo que brinda flexibilidad y escalabilidad para acomodar datos variados y cambiantes.

El modelo lógico desarrollado para esta tarea se diseñó considerando las particularidades de los datos extraídos mediante web scraping. En lugar de tablas con relaciones definidas, este modelo se centró en la organización y almacenamiento de documentos individuales, cada uno de los cuales contenía la información capturada de las páginas web correspondientes. Esto permitió una representación más natural de la estructura de los datos extraídos, evitando la necesidad de definir esquemas rígidos de antemano y facilitando la adaptación a cambios en la estructura de los datos web.

En resumen, MongoDB se utilizó como el sistema de gestión de bases de datos para alojar documentos extraídos a través del web scraping, y se implementó un modelo lógico diseñado específicamente para manejar estos datos no relacionales de manera eficiente y flexible en este entorno. Esto facilitó la organización y el acceso a la información obtenida de manera más efectiva que en un sistema de base de datos relacional tradicional.



Figure 6: Modelo lógico NoSQL

A continuación, se anexan las imágenes de las colecciones de las bases de datos en Mongo DB.



Figure 7: Imagen Base NoSQL MongoDB

6 Automatización de Datos *(Segunda entrega)*

Para desarrollar la Automatización de los datos se realizó en tres apartados distintos que estos son: 1. Araña web, 2. Clusterización y 3. Puntuación, a continuación se expone el avance que tiene cada uno de ellos en el desarrollo del proyecto a la fecha:

1. Araña web: En el proyecto "Brújula Uni", la araña web desempeña un papel esencial en la recolección de información de múltiples fuentes en línea, enfocándose especialmente en la búsqueda y extracción de archivos PDF de los planes de estudio de universidades y programas educativos, adicionalmente los datos recolectados se guardan como datos no estructurados en MongoDB, una base de datos orientada a documentos. Esta estrategia permite manejar la diversidad y complejidad de los datos recolectados de manera eficiente y flexible. Almacenar la información en MongoDB facilita la gestión de grandes volúmenes de datos no estructurados, proporcionando a "Brújula Uni" una base de datos rica y actualizada para apoyar las decisiones educativas de los estudiantes.

CODIGO

```
!pip install selenium
!pip install pymongo
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions
as EC
import requests
from google.colab import files
import os
import shutil
from pymongo import MongoClient
from bson.binary import Binary
from io import BytesIO
import zipfile

Colegio nuestra se ora del rosario

# Configuración del navegador
options = webdriver.ChromeOptions()
options.add_argument('--headless') # Ejecutar en modo
sin cabeza (sin interfaz gráfica)
options.add_argument('--disable-gpu')
options.add_argument('--no-sandbox')
```

```

# Configuración de MongoDB Atlas
uri = "mongodb+srv://malejafan:aleja123@cluster0.ldc8oa.mongodb.net/plandeestudiosrosario?retryWrites=true&w=majority"
client = MongoClient(uri)
db = client['plandeestudiosrosario']

# Lista de URLs de las páginas web que contienen los enlaces a los archivos PDF
urls_paginas_web = [
    'https://urosario.edu.co/licenciatura-en-filosofia ',
    'https://urosario.edu.co/licenciatura-en-ciencias-sociales ',
    # ... Agrega todas las URLs restantes
]

for idx, url_pagina in enumerate(urls_paginas_web, start=1):
    try:
        # Acceder a la página web
        driver = webdriver.Chrome(options=options)
        driver.get(url_pagina)

        # Esperar hasta que el enlace "Descargar plan de estudio" sea visible
        WebDriverWait(driver, 10).until(
            EC.visibility_of_element_located(
                ((By.XPATH, '//a[@class="button" and contains(text(), "Descargar plan de estudio")])')
            )
        )

        # Encontrar el enlace por su clase y texto
        enlace_plan_estudios = driver.find_element(
            By.XPATH, '//a[@class="button" and contains(text(), "Descargar plan de estudio")])')

        # Obtener la URL del enlace
        url_pdf = enlace_plan_estudios.get_attribute('href')

```

```

# Nombre del archivo local para guardar el PDF
nombre_archivo_local =
f'Plan_de_Estudios_{idx}.pdf'

# Realizar la solicitud para descargar el archivo
response = requests.get(url_pdf)

# Verificar si la solicitud fue exitosa
(codigo de estado 200)
if response.status_code == 200:
    # Cargar el contenido del PDF en MongoDB
    db.archivos.insert_one({
        'nombre': nombre_archivo_local,
        'contenido': response.content
    })

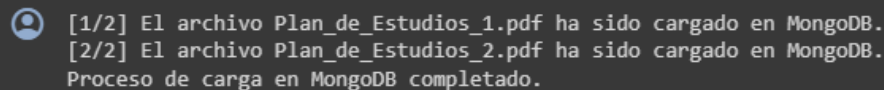
    print(f'[{idx}]/{len(urls_paginas_web)}]
    El archivo {nombre_archivo_local}
    ha sido cargado en MongoDB.')
else:
    print(f'[{idx}]/{len(urls_paginas_web)}]
    Error al descargar el archivo.
    Codigo de estado: {response.status_code}')

except Exception as e:
    print(f'[{idx}]/{len(urls_paginas_web)}]
    Error al procesar la pagina: {str(e)}')
finally:
    # Cerrar el navegador despues de cada iteracion
    driver.quit()

print('Proceso de carga en MongoDB completado.')

```

A continuación se expondrá el ejemplo de la ejecución de la araña web



```

[1/2] El archivo Plan_de_Estudios_1.pdf ha sido cargado en MongoDB.
[2/2] El archivo Plan_de_Estudios_2.pdf ha sido cargado en MongoDB.
Proceso de carga en MongoDB completado.

```

Figure 8: Ejemplo de ejecución de araña web.

2. Clusterización: Una vez recolectados los datos, el proyecto aplicará técnicas de clusterización para organizar esta información de manera coherente y útil. Por ejemplo, se podrían agrupar programas o universidades por áreas de estudio, calidad académica, rango de precios, entre otros factores. Este enfoque ayuda a los usuarios a navegar fácilmente a través de la gran cantidad de información y encontrar programas que se alineen con sus preferencias personales y necesidades académicas. La clusterización es un paso crítico en la creación de una herramienta intuitiva y efectiva que permita comparaciones rápidas y significativas entre diferentes opciones educativas.

3. Puntuación: Finalmente, el proyecto implementará un sistema de puntuación para evaluar y asignar valores a características relevantes de cada programa y universidad. Factores como la ubicación, el costo de la carrera, y la calidad académica, entre otros, serán considerados. Estas puntuaciones permitirán a los usuarios obtener una visión clara y objetiva de cómo cada opción se compara en distintos aspectos clave. La puntuación ayuda a los estudiantes a tomar decisiones educativas más informadas y precisas, proporcionando una evaluación equilibrada y detallada de las opciones disponibles. A continuación, se expondrá el código de implementación de este sistema de puntuación. Cabe aclarar que la automatización de esta sección se realizará mediante una máquina virtual.

CODIGO

```
import pandas as pd

# Carga del archivo Excel
file_path = '/content/datos programas ajustado.xlsx'
data = pd.read_excel(file_path)

# Cálculo de la Calificación de Costo
# -----

# Identificación de la columna de costo
cost_column = 'COSTO.MATRICULA.ESTUD.NUEVOS'

# Manejo de valores faltantes y definición de rangos de costo
costs = data[cost_column].fillna(0)

# Reemplazo de NaN por 0
max_cost = costs.max()
cost_thresholds = [0.25 * max_cost, 0.5 *
max_cost, 0.75 * max_cost, max_cost]
```



```

# Asignacion de calificaciones basadas en umbrales
de costo
cost_ratings = pd.cut(costs, bins=[-1, *cost_thresholds,
float('inf')],
                        labels=[5, 4, 3, 2, 1])

# Calculo de la Calificacion de Ubicacion
# -----

# Para la demostracion, consideramos 'Bogota D.C.'
como una ubicacion deseable
desired_locations = ['Bogota D.C. ']

# Asignacion de calificaciones para la ubicacion
location_ratings =

data['DEPARTAMENTO.OFERTA.PROGRAMA'].
apply(lambda x: 5 if x in desired_locations else 3)

# Calculo de la Calificacion de Periodicidad
# -----

# Columna de periodicidad
periodicity_column = 'PERIODICIDAD.ADMISIONES'

# Suposicion de que las admisiones semestrales
son mas deseables
periodicity_ratings = data[periodicity_column]
.apply(lambda x: 5 if 'Semestral' in x else 3)

# Calculo de la Calificaci n de Acreditacion
# -----

# Suposicion de que 'ESTADO.INSTITUCIN'
indica el estado de acreditacion
accreditation_column = 'ESTADO.INSTITUCIN'

# Suposicion de que
'Activa' indica un estado de acreditacion positivo
accreditation_ratings = data[accreditation_column].apply(
lambda x: 5 if 'Activa' in x else 2)

```

```
# Compilaci n de las calificaciones en un DataFrame
rating_df = pd.DataFrame({
    'Calificaciones de Costo': cost_ratings ,
    'Calificaciones de Ubicaci n ': location_ratings ,
    'Calificaciones de Periodicidad ': periodicity_ratings ,
    'Calificaciones de Acreditaci n ':
    accreditation_ratings
})

# Mostrando las primeras filas de las calificaciones
compiladas
rating_df
```

	Calificaciones de Costo	Calificaciones de Ubicaci3n	Calificaciones de Periodicidad	Calificaciones de Acreditaci3n
1	3	5	4	5
2	5	5	5	3
3	5	4	5	5

Figure 9: Ejemplo de puntuaci3n

Finalmente, se presentar3 un adelanto de c3mo lucir3 la interfaz de usuario a trav3s del front-end del sitio web que planeamos implementar en el futuro. Esta p3gina web, desarrollada utilizando Netlify, se compondr3 de dos secciones , siendo la ultima la mas importante a continuaci3n se expondr3n las dos secciones que se implementaran:



Figure 2: Pag 1. interfaz de usuario.

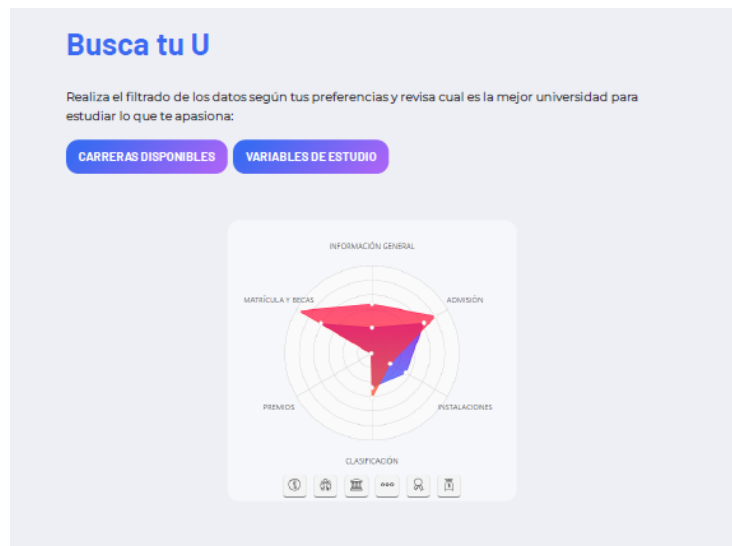


Figure 3: Pag 2. interfaz de usuario.

7 IA (*Segunda entrega*)

En este punto del proyecto, no se lograron desarrollar procesos enfocados al uso de inteligencia artificial en la implementación de la herramienta. Como futura mejora de este proyecto será la implementación de una inteligencia artificial que ofrezca orientación vocacional personalizada al usuario a través de chatbot

8 Proximos pasos (*Tercera entrega*)

A partir de lo desarrollado en este proyecto, se proyecta continuar con la creación de un algoritmo de recomendaciones personalizadas a los usuarios basados en preferencias y necesidades academicas.

Dentro de esto, se realizará una interfaz de usuario para que la herramienta sea intuitiva y facil de usar teniendo en cuenta a los diferentes stakeholders, en esta interfaz se espera implementar funciones de busqueda y filtrado de datos para encontrar tendencias y dar información relevante sobre el desempeño de las instituciones y las carreras en ciencias de la educación.

Finalmente, el desarrollo de este proyecto nos mostrará una herramienta robusta que permita apoyar a los futuros estudiantes de ciencias de la educación en Bogotá dando las herramientas necesarias para tomar una decisión informada con respecto a la institución educativa

Para optimizar la automatización del proyecto, los siguientes pasos involucran la implementación de máquinas virtuales. Estas máquinas se encargarán de ejecutar la automatización de procesos clave como la clusterización, la puntuación y la extracción de información mediante la técnica de araña web. Además, se planea desarrollar el back-end del sistema, un componente crucial para el procesamiento y gestión de los datos recolectados. Finalmente, se procederá con el diseño y montaje de la interfaz de usuario a través de una página web, lo que facilitará la interacción de los usuarios con la plataforma y mejorará significativamente la experiencia del usuario.

9 Lecciones aprendidas *(Tercera entrega)*

Aprendimos a desarrollar una araña web eficiente, una herramienta clave para descargar y analizar información como los planes de estudios de diferentes universidades. Esta experiencia nos enseñó la importancia de los Términos de Servicio de los sitios web y la particularidad que cada uno de estos tiene. También descubrimos cómo manejar datos no estructurados con MONGODB y la importancia de estructurar adecuadamente los datos recolectados, lo que requiere habilidades en el manejo de bases de datos. Además, reforzamos nuestras capacidades en el manejo de errores y excepciones, asegurando así la eficacia y continuidad de nuestros procesos de rastreo.

En el diseño de diagramas de flujo de datos, aprendimos la importancia de la claridad y la coherencia en la representación de la información. Comprendimos que una simbología consistente y una documentación detallada de cada paso son cruciales para la comprensión y el seguimiento eficaz del flujo de datos. Identificar y explicar claramente los puntos de decisión se convirtió en una práctica esencial. Esta experiencia también destacó la importancia de diseñar diagramas flexibles y escalables, preparados para adaptarse a los cambios y facilitar su mantenimiento a largo plazo.

En la automatización e integración de datos, aprendimos a establecer objetivos claros que orientan el desarrollo y la selección de herramientas adecuadas para nuestros propósitos. Comprendimos la importancia de integrar la seguridad y la privacidad de los datos desde el principio del proyecto. Las pruebas se revelaron como un paso fundamental para asegurar la funcionalidad de nuestros procesos automatizados. Por último, la implementación de un sistema de monitoreo y mantenimiento continuo nos enseñó la importancia de estar preparados para detectar y solucionar problemas rápidamente, manteniendo el sistema optimizado y actualizado.

10 Bibliografía

Cardenas, S. (03 de agosto de 2015). www.elcolombiano.com. Recuperado el 15 de julio de 2016, de www.elcolombiano.com:

<https://www.elcolombiano.com/colombia/educacion/estas-son-las-carreras-que-menos-estudian-los-colombianos-MA2471952>

López Fernández, M., Sánchez Herrera, S. (2018). Relación entre la madurez vocacional y la motivación hacia el aprendizaje académico. *International Journal of Developmental and Educational Psychology*, 1(1), 21-30.

SPADIES - Estadísticas de deserción. (2023). SPADIES. Recuperado el 25 de noviembre de 2023, de

<https://www.mineducacion.gov.co/sistemasinfo/spadies/secciones/Estadisticas-de-desercion/>

Tintaya Condori, P. (2016). Orientación profesional y satisfacción profesional. *Reflexiones en psicología*, (15), 45-58.