Aplicación de Modelos Mixtos en Data Analytics para el Análisis Comparativo de Resultados de las Pruebas Saber 11 y Saber Pro en Colombia

Edwin Castiblanco[1], Rodrigo Zambrano[2]
[1−2]Departamento de Ciencias, Universidad Central
Maestría en Ciencia de Datos
Curso de Bases de Datos
Bogotá, Colombia
ecastiblancos@ucentral.edu.co[1], rzambranol1@ucentral.edu.co[2]

29 de noviembre de 2024

Índice

1. Introducción

En el contexto educativo colombiano, las Pruebas Saber 11 y Saber Pro son evaluaciones para medir el rendimiento de los estudiantes en diversas etapas de su formación. Sin embargo, interpretar los resultados y su relación con variables como la edad, el lugar de residencia y el puntaje obtenido en las pruebas Saber 11 del 2019 sigue siendo un reto. Abordar este problema es clave, ya que la falta de análisis adecuado puede llevar a interpretaciones erróneas y decisiones de política menos acertadas, en este sentido, buscando normalizar la base de datos se hizo una unión entre bases de datos por el código único del colegio de donde los estudiantes se graduaron y se agrupan los datos por instituciones de educación superior.

Es necesario aplicar modelos mixtos en el análisis de datos para entender cómo estas variables influyen en los resultados de las pruebas Saber Pro. Esto es relevante en Colombia, donde la diversidad geográfica y socioeconómica puede afectar el rendimiento de los estudiantes y las instituciones (García y Martínez, ,). Así mismo, el promedio de los índices de rendimiento en colegios . .

El uso de técnicas avanzadas de análisis de datos permitirá obtener una visión más clara de los factores que influyen en los resultados. Los modelos mixtos, que permiten incorporar efectos fijos y aleatorios, son útiles para desentrañar las interacciones entre las variables mencionadas. Esta metodología puede mejorar la comprensión de cómo el contexto socioeconómico y educativo impacta en los resultados, lo que podría orientar la formulación de políticas más justas y efectivas(López y Gómez, ,).

2. Características del proyecto de investigación que hace uso de Bases de Datos

El uso de bases de datos en la investigación educativa, y particularmente en el análisis de las pruebas Saber Pro en Colombia, ofrece múltiples ventajas que permiten la profundización en el análisis de grandes volúmenes de información. Este proyecto se sustenta en el acceso y manejo de una base de datos robusta que contiene información detallada sobre el desempeño académico de los estudiantes en las pruebas Saber Pro, lo cual permite no solo identificar patrones generales, sino también explorar relaciones específicas entre variables como el entorno socioeconómico, el rendimiento por áreas de conocimiento y otros factores relevantes.

Una de las principales características de este tipo de investigación es la objetividad en la recolección y procesamiento de los datos. A diferencia de estudios cualitativos que se basan en encuestas o entrevistas, la investigación con bases de datos extrae información ya consolidada, eliminando sesgos subjetivos en la recopilación de datos. Este enfoque permite trabajar con una muestra amplia y representativa de la población estudiantil, lo cual incrementa la validez externa de los resultados obtenidos.

Otra característica clave es la capacidad de análisis longitudinal, que permite observar tendencias a lo largo del tiempo. En el caso de la base de datos de las pruebas Saber Pro, se puede evaluar cómo ha cambiado el desempeño estudiantil en diferentes áreas a lo largo de los años, qué factores han influido en dichos cambios, y cómo estas variaciones se relacionan con modificaciones en el sistema educativo.

2.1. Modelos Mixtos en Data Analytics para el Análisis Comparativo de Pruebas Saber 11 y Pro en Colombia (2017-2021)

Este estudio busca desarrollar un análisis detallado de los resultados de las pruebas Saber 11 y Saber Pro en Colombia, utilizando modelos mixtos en analítica de datos. Se evaluará cómo variables contextuales como la edad, el contexto de residencia y el puntaje en Saber 11 de 2019 influyen en el rendimiento de los estudiantes en Saber Pro. Los modelos mixtos permitirán examinar estas relaciones de manera más precisa. A partir de estos hallazgos, se proporcionarán recomendaciones que apoyen la toma de decisiones en políticas educativas, mejorando así el sistema educativo en Colombia.

2.2. Objetivo general

Determinar cuáles son los posibles factores que inciden en los resultados obtenidos por las instituciones en el puntaje global de la prueba Saber-Pro de 2019.

2.2.1. Objetivos especificos

- Caracterizar los patrones y los efectos de las variaciones en los resultados académicos de LAS INSTITUCIONES.
- Determinar el impacto de las variables contextuales en el desempeño académico de acuerdo a la variable objetivo que es el puntaje global.

2.3. Alcance

La presente investigación se enfoca en analizar los datos de las Pruebas Saber 11 y Saber Pro en Colombia, correspondientes al año 2021. algunas variables que se considerarán son: la edad de los estudiantes, el contexto de residencia (urbano o rural), y el puntaje obtenido en las pruebas Saber 11. La población de estudio está constituida por los estudiantes que participaron en ambas pruebas durante el año 2019, seleccionando una muestra representativa a nivel nacional. Este enfoque permite un análisis detallado de cómo estas variables contextuales pueden influir en los resultados obtenidos en las Pruebas Saber Pro, facilitando la identificación de patrones y discrepancias en el rendimiento académico en función de factores socioeconómicos y geográficos. La elección de este periodo temporal se basa en la disponibilidad de datos actualizados y completos para el año 2019, lo cual es esencial para obtener resultados relevantes y precisos.

2.4. Pregunta de investigación

¿Qué factores inciden en el puntaje global de los estudiantes que presentan las pruebas Saber-Pro?.

2.5. Hipótesis

El rendimiento de los estudiantes en las Pruebas Saber Pro está significativamente influenciado por variables contextuales como la edad, el lugar de residencia y el puntaje obtenido en las Pruebas Saber 11. La aplicación de modelos mixtos permitirá identificar cómo estas variables, en combinación con la diversidad geográfica y socioeconómica, afectan el desempeño académico de manera diferenciada, lo que podría contribuir a un diseño más equitativo de las políticas educativas en Colombia.

3. Reflexiones sobre el origen de datos e información

El valor de la objetividad en la investigación basada en datos: El uso de bases de datos robustas para el análisis educativo, como en el caso de las pruebas Saber Pro, permite un enfoque más objetivo y cuantitativo en comparación con estudios cualitativos. Al trabajar con información consolidada y no depender de percepciones subjetivas de los participantes, se minimizan los sesgos, lo que refuerza la fiabilidad de los resultados. Esto es crucial en la formulación de políticas educativas, ya que los datos empíricos permiten una interpretación más precisa de la realidad educativa.

La importancia de una muestra representativa para la validez externa: La capacidad de trabajar con una base de datos que abarca una amplia muestra de la población estudiantil es clave para garantizar que los resultados obtenidos sean generalizables. Este tipo de análisis permite extraer conclusiones que pueden aplicarse a nivel nacional, lo que es vital para entender patrones de desempeño académico y diseñar intervenciones educativas que respondan a las necesidades de un amplio espectro de estudiantes. La validez externa es un componente fundamental para que los hallazgos puedan influir positivamente en las políticas educativas.

El análisis longitudinal como herramienta para detectar tendencias y cambios educativos: La posibilidad de realizar análisis longitudinales en bases de datos educativas ofrece un panorama más profundo sobre cómo evoluciona el rendimiento académico en distintas áreas a lo largo del tiempo. Este enfoque es esencial para evaluar los efectos de cambios en el sistema educativo o en el contexto socioeconómico del país. Al detectar tendencias y factores que influyen en el desempeño de los estudiantes, los investigadores pueden sugerir mejoras en las políticas educativas que respondan a problemas identificados y optimicen los resultados futuros.

3.1. ¿Cuál es el origen de los datos e información?

Se cuenta con dos sets de datos los cuales representan el primero los resultados por instituciones educativas de educación media en la pruebas Saber 11 y el segundo presenta los resultados de la pruebas Saber-Pro correspondiente a las instituciones de educación superior desde el año 2017 hasta el año 2021 para cada caso.

3.2. ¿Cuáles son las consideraciones legales o eticas del uso de la información?

Dando inicio a lo que conlleva desde la parte legal y ética de la protección de la información de los datos esta dada desde el artículo 11 de la CIDH se plantea como primer paso la protección de la información (Petrino,).

En contraparte la primera ley promulgada sobre la protección y resguardo de los datos que se le hacen a las personas cuando llenan formularios de forma electrónica esta ampara por la **Ley 1581 de 2012** (Ronderos,) presentando consideraciones tales como:

La institución receptora de la información no podra hacer uso diferente de los datos al consentimiento informado al usuario.

No se puede vender o suministrar los datos a terceros sin la previa autorización del dueño primario de los datos.

3.3. ¿Cuáles son los retos de la información y los datos que utilizara en la base de datos en terminos de la calidad y la consolidación?

la calidad de los datos es fundamental, ya que la presencia de datos incompletos, duplicados o inconsistentes puede afectar la integridad del análisis. Es esencial establecer un proceso de limpieza de datos para detectar y corregir errores, así como asegurar que los datos sean precisos y coherentes. Otro desafío importante es la consolidación de la información, especialmente cuando los datos provienen de múltiples fuentes. La integración de estos conjuntos puede generar problemas si no se aplican formatos estandarizados o si existen diferencias en la estructura de los datos.

3.4. ¿Qué espera de la utilización de un sistema de Bases de Datos para su proyecto?

Se busca mejorar significativamente la gestión y organización de la información, permitiendo un acceso eficiente y estructurado a los datos. Espero que este sistema facilite el almacenamiento seguro y centralizado de grandes volúmenes de información, evitando problemas como la duplicación o pérdida de datos. Además, confío en que la base de datos permitirá realizar consultas rápidas y precisas, optimizando el tiempo invertido en la búsqueda y el análisis de información relevante para el proyecto.

4. Diseño del Modelo de Datos

- **Caracterización**: Variables como fecha de nacimiento, género, departamento de residencia, entre otras.
- **Colegio**: Código DANE del colegio, caracterización de la institución de Saber 11.
- **Hogar**: Caracterización del hogar del estudiante, como acceso a internet, educación de los padres.
- **Examen**: Caracterización de la prueba Saber-Pro por cada estudiante.
- **IES**: Caracterización de la universidad donde se presentó el examen Saber-Pro.

4.1. Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto (*Primera entrega*)

* SQL Developer es una herramienta poderosa proporcionada por Oracle para el desarrollo y administración de bases de datos SQL. Sus principales funciones incluyen la creación, modificación y gestión de esquemas de bases de datos, así como la ejecución de consultas SQL y la administración de objetos de base de datos como tablas, índices, vistas y procedimientos almacenados. SQL Developer permite a los desarrolladores escribir, ejecutar y depurar sentencias SQL y PL/SQL, facilitando el análisis y la optimización de consultas a través de su planificador de ejecución. Además, ofrece capacidades para la migración de bases de datos de otros sistemas hacia Oracle, y es una herramienta fundamental para la gestión de usuarios, roles y permisos, lo que contribuye a la seguridad del entorno de datos (Gupta,).

* Oracle SQL Data Modeler, por su parte, es una herramienta enfocada en el diseño y modelado de bases de datos, permitiendo la creación de modelos conceptuales, lógicos y físicos. Data Modeler facilita la representación visual de la estructura de la base de datos mediante diagramas entidad-relación (ER), lo que simplifica la creación y modificación del diseño de datos antes de su implementación en el sistema de gestión de bases de datos. Entre sus funciones destacadas están la generación automática de esquemas SQL a partir de los modelos diseñados, la ingeniería inversa para visualizar bases de datos existentes y la verificación de consistencia y normalización de los datos. Además, ayuda a los desarrolladores y administradores de bases de datos a optimizar el diseño de las tablas y relaciones, garantizando la integridad y eficiencia del sistema de gestión de datos (Bock y Yager,).

4.2. Diagrama modelo de datos

EL diagrama modelo presenta las entidades relación de las principales etiquetas que fueron diseñadas con el fin de poder hacer busquedas eficientes y sencillas dentro del modelo.

4.3. Imágenes de la Base de Datos

A continuación en la figura 2 el modelo ER para el tratamiento de los datos.

ESTU_DEPTO_RESIDE	ESTU_COD_RESIDE_DEPTO	ESTU_MCPIO_RESIDE	ESTU_COD_RESIDE_MCPIO	ESTU_COLE_TERMINO	Código Dane	ESTU_COD_COLE_MCPIO_TERMINO	ESTU_TITULOOBTENIDOBACHILLER	ESTU_VALORMATRICULAUEXT	ESTU_VALORMATRICULAUNIVERSIDAD
MAGDALENA	47.0	SANTA MARTA	47001.0	COL NACIONAL MIXO DE BACHILLERATO	147245000000.0	47245.0	Bachiller técnico	Mas de 7 millones	Menos de 500 mil
MAGDALENA	47.0	SANTA MARTA	47001.0	ESCUELA NORMAL SAN PEDRO ALEJANDRINO	247001000000.0	47001.0	Bachiller académico	NaN	Menos de 500 mil
MAGDALENA	47.0	SANTA MARTA	47001.0	INSTITUCION EDUCATIVA DISTRITAL TECNICO INDUST...	147001000000.0	47001.0	Bachiller técnico	NaN	Entre 500 mil y menos de 1 millón
MAGDALENA	47.0	CIÉNAGA	47189.0	IE SAN JUAN DEL CORDOBA	147189000000.0	47189.0	Bachiller académico	NaN	Entre 1 millón y menos de 2.5 millones
MAGDALENA	47.0	SANTA MARTA	47001.0	IED JACQUELINE KENNEDY	147001000000.0	47001.0	Bachiller académico	NaN	Menos de 500 mil

Figura 1: Sistema de variables

4.4. Código SQL - lenguaje de definición de datos (DDL)

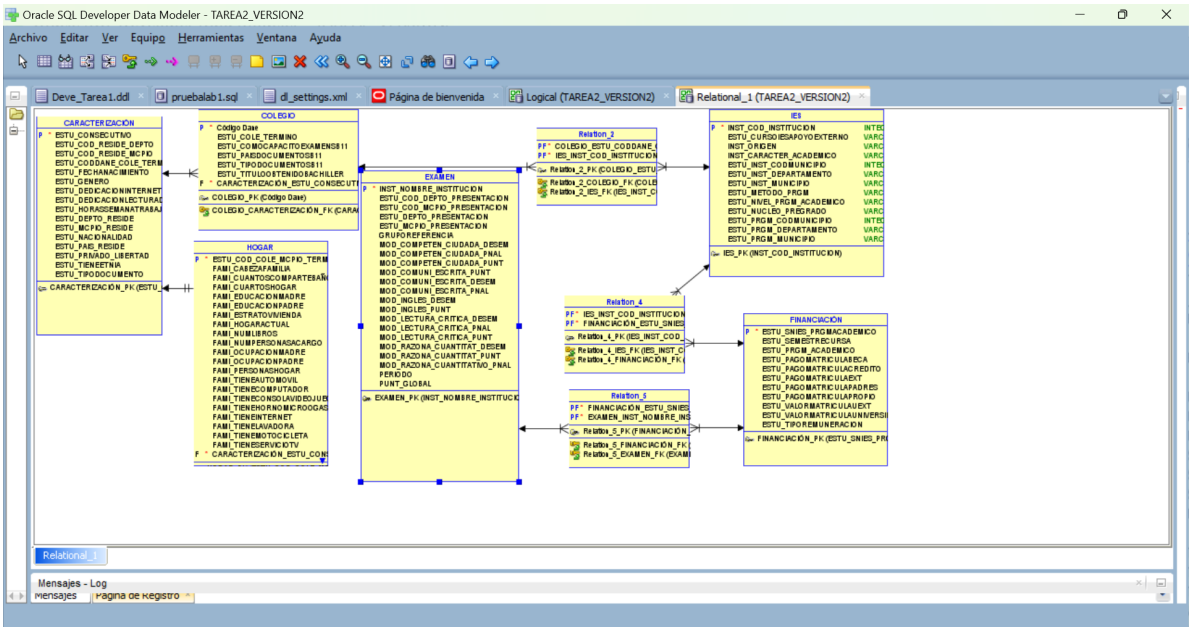

Figura 2: Sistema DDL base de datos

4.5. Código SQL - Manipulación de datos (DML)

```sql
CREATE TABLE caracterizaci n (
    estu_consecutivo              INTEGER NOT NULL,
    estu_cod_reside_depto         INTEGER,
    estu_cod_reside_mcpio         INTEGER,
    estu_coddane_cole_termino     INTEGER,
    estu_fechanacimiento          DATE,
    estu_genero                   CHAR(1),
    estu_dedicacioninternet       INTEGER,
    estu_dedicacionlecturadiaria  INTEGER,
    estu_horassemanatrabaja       INTEGER,
    estu_depto_reside             VARCHAR2(50 CHAR),
    estu_mcpio_reside             VARCHAR2(70 CHAR),
    estu_nacionalidad             VARCHAR2(50 CHAR),
    estu_pais_reside              VARCHAR2(100 CHAR),
    estu_privado_libertad         CHAR(1),
    estu_tieneetnia               CHAR(1),
    estu_tipodocumento            VARCHAR2(45 CHAR)
);

ALTER TABLE caracterizaci n ADD CONSTRAINT caracterizaci n_pk PRIMARY KEY (
    estu_consecutivo);

CREATE TABLE colegio (
    "C digo Dane"                  INTEGER NOT NULL,
    estu_cole_termino             VARCHAR2(150 CHAR),
    estu_comocapacitoexamensb11   VARCHAR2(300 CHAR),
    estu_paisdocumentosb11        VARCHAR2(100 CHAR),
    estu_tipodocumentosb11        VARCHAR2(50 CHAR),
    estu_tituloobtenidobachiller  VARCHAR2(200 CHAR),
```

```
    caracterizaci n_estu_consecutivo INTEGER NOT NULL
);

ALTER TABLE colegio ADD CONSTRAINT colegio_pk PRIMARY KEY ("C digo Dane");

CREATE TABLE examen (
    inst_nombre_institucion      VARCHAR2 NOT NULL,
    estu_cod_depto_presentacion  INTEGER,
    estu_cod_mcpio_presentacion  INTEGER,
    estu_depto_presentacion      VARCHAR2(50 CHAR),
    estu_mcpio_presentacion      VARCHAR2(70 CHAR),
    gruporeferencia              VARCHAR2(100 CHAR),
    mod_competen_ciudada_desem   INTEGER,
    mod_competen_ciudada_pnal    INTEGER,
    mod_competen_ciudada_punt    INTEGER,
    mod_comuni_escrita_punt      INTEGER,
    mod_comuni_escrita_desem     INTEGER,
    mod_comuni_escrita_pnal      INTEGER,
    mod_ingles_desem             INTEGER,
    mod_ingles_punt              INTEGER,
    mod_lectura_critica_desem    INTEGER,
    mod_lectura_critica_pnal     INTEGER,
    mod_lectura_critica_punt     INTEGER,
    mod_razona_cuantitat_desem   INTEGER,
    mod_razona_cuantitat_punt    INTEGER,
    mod_razona_cuantitativo_pnal INTEGER,
    periodo                      INTEGER,
    punt_global                  INTEGER
);

ALTER TABLE examen ADD CONSTRAINT examen_pk PRIMARY KEY (inst_nombre_institucion);

CREATE TABLE financiaci n (
    estu_snies_prgmacademico       INTEGER NOT NULL,
    estu_semestrecursa             INTEGER,
    estu_prgm_academico            VARCHAR2,
    estu_pagomatriculabeca         CHAR(1),
    estu_pagomatriculacredito      CHAR(1),
    estu_pagomatriculaext          CHAR(1),
    estu_pagomatriculapadres       CHAR(1),
    estu_pagomatriculapropio       CHAR(1),
    estu_valormatriculauext        INTEGER,
    estu_valormatriculauniversidad INTEGER,
    estu_tiporemuneracion          VARCHAR2(200 CHAR)
);

ALTER TABLE financiaci n ADD CONSTRAINT financiaci n_pk PRIMARY KEY (
    estu_snies_prgmacademico);

CREATE TABLE hogar (
    estu_cod_cole_mcpio_termino      INTEGER NOT NULL,
    fami_cabezafamilia               CHAR(1),
    fami_cuantoscomparteba o          INTEGER,
    fami_cuartoshogar                INTEGER,
    fami_educacionmadre              VARCHAR2(150 CHAR),
    fami_educacionpadre              VARCHAR2(150 CHAR),
    fami_estratovivienda             INTEGER,
    fami_hogaractual                 VARCHAR2(150 CHAR),
    fami_numlibros                   INTEGER,
```

```
87      fami_numpersonasacargo            INTEGER,
88      fami_ocupacionmadre               VARCHAR2(150 CHAR),
89      fami_ocupacionpadre               VARCHAR2(150 CHAR),
90      fami_personashogar                INTEGER,
91      fami_tieneautomovil               CHAR(1),
92      fami_tienecomputador              CHAR(1),
93      fami_tieneconsolavideojuegos      CHAR(1),
94      fami_tienehornomicroogas          CHAR(1),
95      fami_tieneinternet                CHAR(1),
96      fami_tienelavadora                CHAR(1),
97      fami_tienemotocicleta             CHAR(1),
98      fami_tieneserviciotv              CHAR(1),
99      caracterizaci n_estu_consecutivo INTEGER NOT NULL
100 );
101
102 CREATE UNIQUE INDEX hogar__idx ON hogar (caracterizaci n_estu_consecutivo ASC);
103
104 ALTER TABLE hogar ADD CONSTRAINT hogar_pk PRIMARY KEY (estu_cod_cole_mcpio_termino
        );
```

4.6. Código SQL + Resultados: Vistas

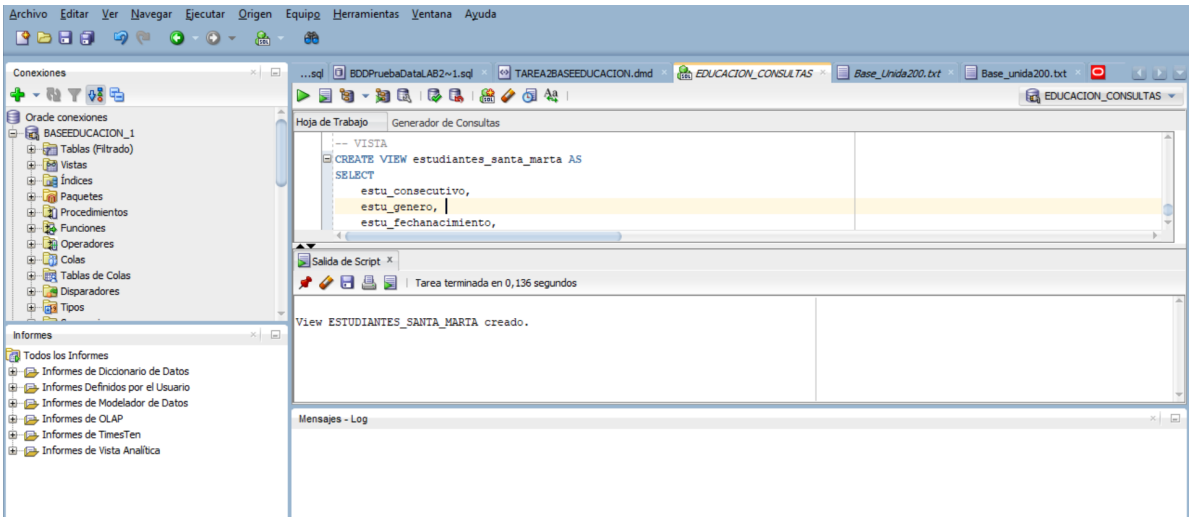

Figura 3: Vista de la base

```
CREATE VIEW estudiantes_santa_marta AS
SELECT
    estu_consecutivo,
    estu_genero,
    estu_fechanacimiento,
    estu_depto_reside,
    estu_mcpio_reside,
    estu_nacionalidad
FROM caracterización
WHERE estu_mcpio_reside = 'SANTA MARTA';
```

4.7. Código SQL + Resultados: Triggers

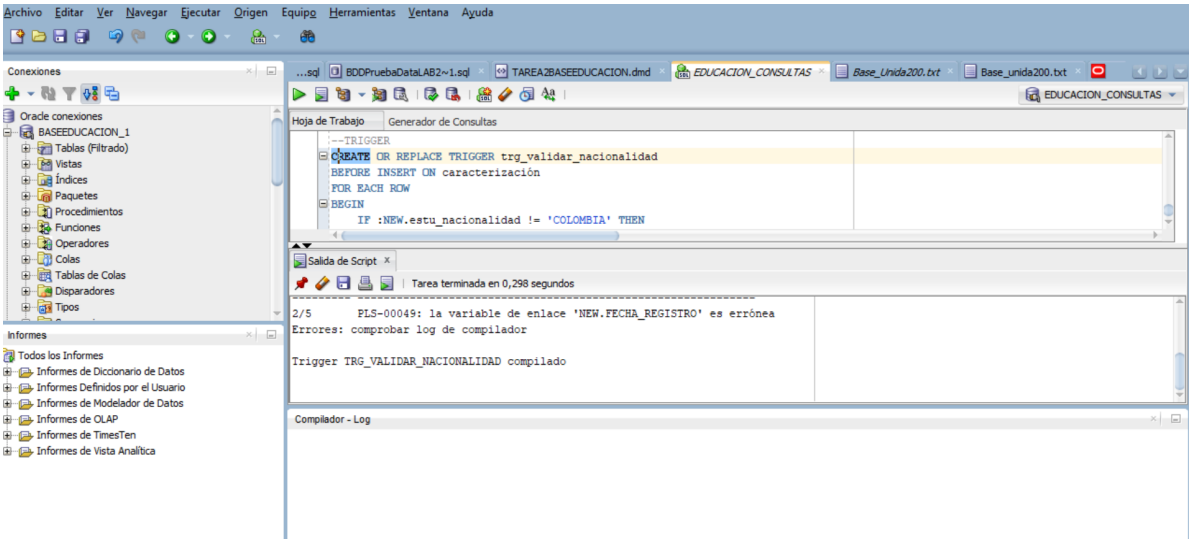

Figura 4: Trigger

–TRIGGER article listings xcolor

```
CREATE OR REPLACE TRIGGER trg_validar_nacionalidad

BEFORE INSERT ON caracterizaci n
FOR EACH ROW
BEGIN
    IF :NEW.estu_nacionalidad != 'COLOMBIA' THEN
        RAISE_APPLICATION_ERROR(-20001, 'Solo se permiten estudiantes de nacionalidad
            COLOMBIANA');
    END IF;
END;
```

4.8. Código SQL + Resultados: Funciones

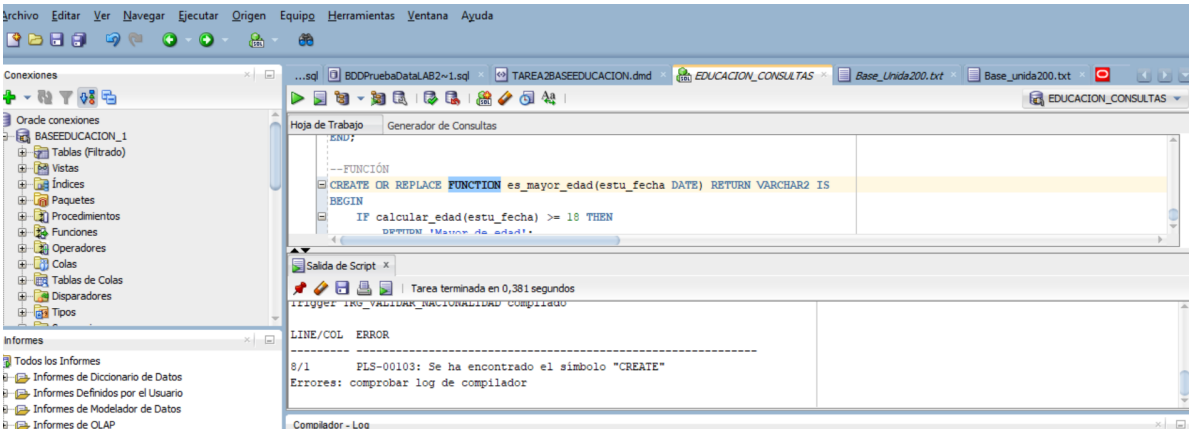

Figura 5: Función es_mayor_edad

Listing 1: Función SQL para determinar si el estudiante es mayor de edad

```sql
CREATE OR REPLACE FUNCTION es_mayor_edad(estu_fecha DATE)
RETURN VARCHAR2 IS
BEGIN
    IF calcular_edad(estu_fecha) >= 18 THEN
        RETURN 'Mayor de edad';
    ELSE
        RETURN 'Menor de edad';
    END IF;
END;
```

4.9. Código SQL + Resultados: Procedimientos Almacenados

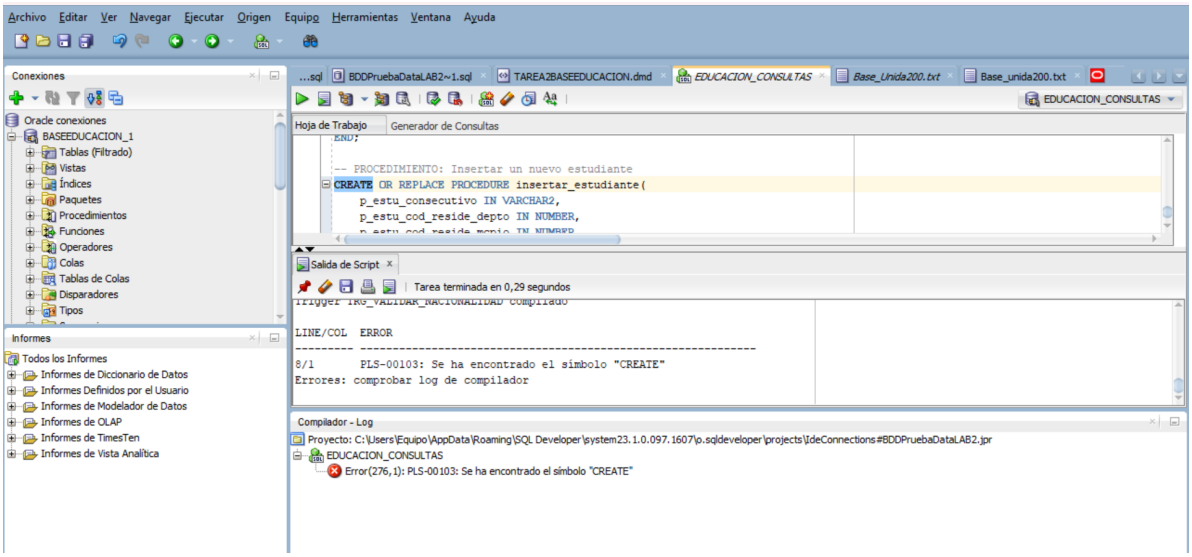

Figura 6: Procedimiento Almacenado para Insertar Estudiante

Listing 2: Procedimiento almacenado para insertar un nuevo estudiante

```
CREATE OR REPLACE PROCEDURE insertar_estudiante(
    p_estu_consecutivo IN VARCHAR2,
    p_estu_cod_reside_depto IN NUMBER,
    p_estu_cod_reside_mcpio IN NUMBER,
    p_estu_coddane_cole_termino IN NUMBER,
    p_estu_fechanacimiento IN DATE,
    p_estu_genero IN CHAR,
    p_estu_dedicacioninternet IN VARCHAR2,
    p_estu_dedicacionlecturadiaria IN VARCHAR2,
    p_estu_horassemanatrabaja IN NUMBER,
    p_estu_depto_reside IN VARCHAR2,
    p_estu_mcpio_reside IN VARCHAR2,
    p_estu_nacionalidad IN VARCHAR2
) IS
BEGIN
    INSERT INTO caracterizaci n (estu_consecutivo, estu_cod_reside_depto,
        estu_cod_reside_mcpio, estu_coddane_cole_termino, estu_fechanacimiento, estu_genero,
         estu_dedicacioninternet, estu_dedicacionlecturadiaria, estu_horassemanatrabaja,
        estu_depto_reside, estu_mcpio_reside, estu_nacionalidad)
    VALUES (p_estu_consecutivo, p_estu_cod_reside_depto, p_estu_cod_reside_mcpio,
        p_estu_coddane_cole_termino, p_estu_fechanacimiento, p_estu_genero,
        p_estu_dedicacioninternet, p_estu_dedicacionlecturadiaria, p_estu_horassemanatrabaja
        , p_estu_depto_reside, p_estu_mcpio_reside, p_estu_nacionalidad);
END;
```

5. Bases de Datos No-SQL

Las bases de datos NoSQL (del inglés *Not Only SQL*) son un conjunto de tecnologías diseñadas para almacenar, gestionar y consultar datos que no se ajustan al modelo relacional tradicional. A diferencia de las bases de datos SQL, las NoSQL están optimizadas para manejar grandes volúmenes de datos distribuidos en tiempo real, siendo especialmente útiles para aplicaciones modernas como análisis de datos, redes sociales y procesamiento de datos de sensores.

Existen diferentes tipos de bases de datos NoSQL, cada una adecuada para casos de uso específicos:

- **Bases de datos clave-valor:** Usan un modelo simple de pares clave-valor, ideal para datos de lectura y escritura rápida (*e.g., Redis*).
- **Bases de datos de documentos:** Almacenan datos en formato JSON o BSON, permitiendo estructuras flexibles y jerárquicas (*e.g., MongoDB*).
- **Bases de datos en columna:** Optimizadas para grandes cantidades de datos tabulares, como en análisis de Big Data (*e.g., Cassandra*).
- **Bases de datos de grafos:** Diseñadas para modelar relaciones complejas entre datos (*e.g., Neo4j*).

El uso de bases NoSQL ha crecido significativamente debido a su capacidad para manejar datos no estructurados y su escalabilidad horizontal, características esenciales en el desarrollo de aplicaciones modernas (Strauch, ,).

5.1. Diagrama Bases de Datos No-SQL

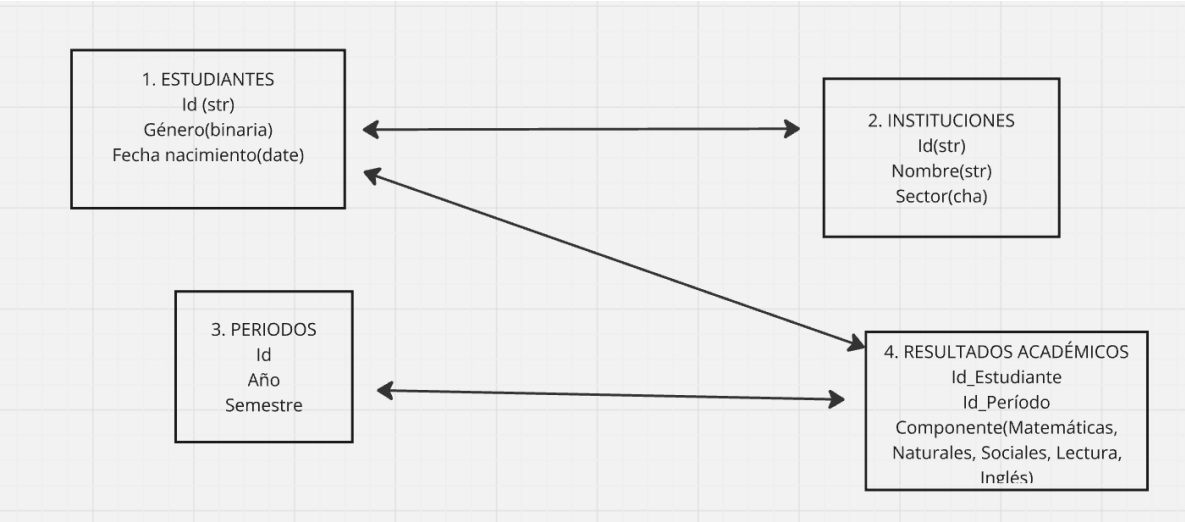

Figura 7: NoSQL

De acuerdo a la imagén presentada se tiene:

Explicación de la Base de Datos NoSQL

En esta sección, se describe la estructura de la base de datos NoSQL basada en el diagrama proporcionado. Esta base está organizada en cuatro colecciones principales: **Estudiantes**, **Instituciones**, **Periodos** y **Resultados Académicos**, con relaciones entre ellas para modelar la información.

1. Colección: Estudiantes

La colección **Estudiantes** contiene los datos básicos de los estudiantes. Cada documento incluye los siguientes campos:

- **Id**: Identificador único del estudiante (`string`).
- **Género**: Representa el género del estudiante como un valor binario.
- **Fecha de nacimiento**: Fecha de nacimiento del estudiante (`date`).

Ejemplo de documento:

```
{
  "id": "E001",
  "genero": "M",
  "fecha_nacimiento": "2005-03-15"
}
```

2. Colección: Instituciones

Esta colección almacena las instituciones educativas asociadas con los estudiantes. Los campos principales son:

- **Id**: Identificador único de la institución (`string`).
- **Nombre**: Nombre de la institución (`string`).
- **Sector**: Tipo de sector, como público o privado (`char`).

Ejemplo de documento:

```
{
  "id": "I001",
  "nombre": "Instituto Educativo Central",
  "sector": "Público"
}
```

3. Colección: Periodos

La colección **Periodos** describe los periodos académicos. Contiene los siguientes campos:

- **Id**: Identificador único del periodo (`string`).
- **Año**: Año del periodo (`int`).
- **Semestre**: Semestre correspondiente (`int`).

Ejemplo de documento:

```
{
  "id": "P001",
  "año": 2024,
  "semestre": 1
}
```

4. Colección: Resultados Académicos

La colección **Resultados Académicos** almacena los resultados de los estudiantes en diferentes componentes académicos. Los campos son:

- **Id_Estudiante**: Referencia al `Id` del estudiante en la colección **Estudiantes**.
- **Id_Periodo**: Referencia al `Id` del periodo en la colección **Periodos**.
- **Componente**: Un objeto que incluye los resultados en Matemáticas, Naturales, Sociales, Lectura e Inglés.

Ejemplo de documento:

```
{
  "id_estudiante": "E001",
  "id_periodo": "P001",
  "componentes": {
    "matematicas": 85,
    "naturales": 90,
    "sociales": 80,
    "lectura": 75,
    "ingles": 88
  }
}
```

Relaciones entre Colecciones

1. **Estudiantes - Resultados Académicos:** La colección **Resultados Académicos** referencia el `Id` de los estudiantes para asociar los resultados con cada persona.
2. **Periodos - Resultados Académicos:** Cada registro de resultados académicos está asociado a un periodo específico mediante el `Id_Periodo`.
3. **Estudiantes - Instituciones:** Aunque no está explícitamente en la estructura, puede añadirse una referencia en **Estudiantes** para indicar a qué institución pertenece cada estudiante.

Modelo Optimizado

En un modelo optimizado para consultas frecuentes, se pueden combinar datos de instituciones y periodos en la colección **Resultados Académicos**.

Ejemplo de documento combinado:

```
{
  "id_estudiante": "E001",
  "id_periodo": "P001",
  "periodo": {
    "año": 2024,
    "semestre": 1
  },
  "institucion": {
    "id": "I001",
    "nombre": "Instituto Educativo Central",
    "sector": "Público"
  },
  "componentes": {
    "matematicas": 85,
    "naturales": 90,
```

```
    "sociales": 80,
    "lectura": 75,
    "ingles": 88
  }
}
```

Este diseño minimiza consultas cruzadas y mejora el rendimiento en bases de datos NoSQL.

5.2. SMBD utilizado para la Base de Datos No-SQL (*Segunda entrega*)

Para realizar el ejercicio de la base de datos No SQL utilizamos el programa MONGO DB en su servicio Atlas que es el que permite trabajar con bases de datos no relacinales en la nube.

MongoDB es una base de datos NoSQL orientada a documentos que se ha vuelto muy popular en el desarrollo de aplicaciones modernas. A diferencia de las bases de datos relacionales tradicionales, MongoDB almacena la información en documentos flexibles tipo JSON, lo que permite una estructura de datos más dinámica y escalable. Es útil en proyectos de desarrollo web, aplicaciones móviles, gestión de contenidos, análisis en tiempo real y sistemas que requieren alta disponibilidad y escalamiento horizontal.

MongoDB ofrece ventajas significativas en entornos empresariales donde se manejan grandes volúmenes de datos no estructurados o semi-estructurados. Sus principales características incluyen la posibilidad de manejar esquemas dinámicos, consultas rápidas, indexación flexible y capacidad para replicación y fragmentación de datos. Empresas como Google, Facebook, Nokia y eBay utilizan MongoDB para gestionar sus enormes volúmenes de información.

Utilizamos una muestra de 199 datos de nuestra base de datos de los resultados de la prueba Saber pro 2021, y generamos un codigo en python para conectar nuestra muestra a la plataforma de MONGO en la nube y a continuación se encuentra el enlace al cuaderno de collab para verificar el codigo y la evidencia de la conexión.

Acceder al Código en Google Colab

5.3. Diagrama Bases de Datos No-SQL

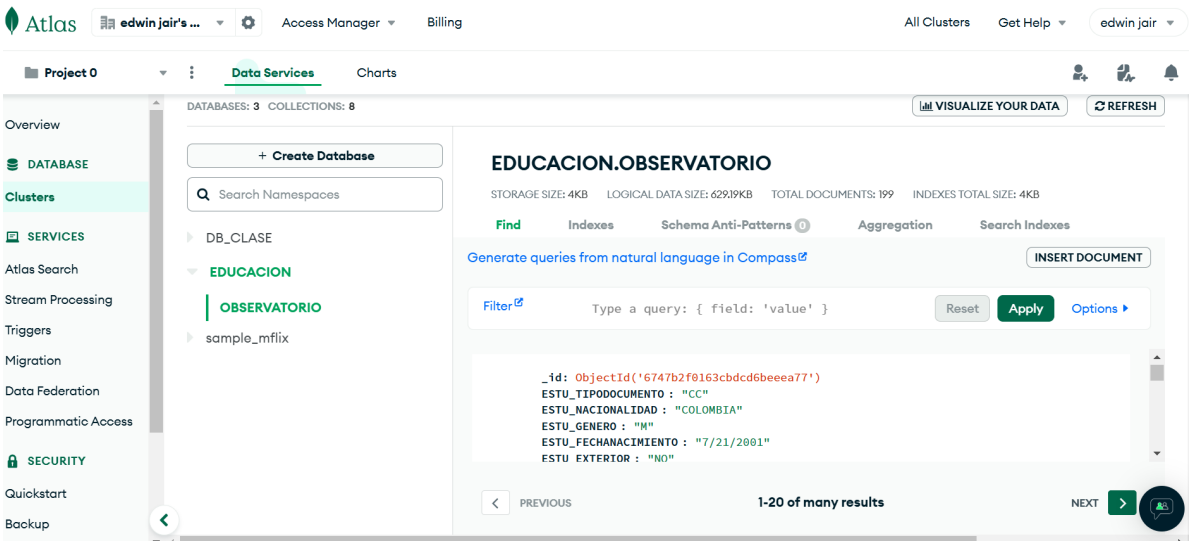

Figura 8: Base NoSQL de Saber Pro 2021

6. Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos (*Tercera entrega*)

Para el ETL de la base de datos utilizamos el software KNIME que nos permite extraer información, cargarla en una plataforma de edición, comibanarla o transformarla en los requerimientos especificos y cargarla o publicarla para su respectivo consumo.

KNIME (Konstanz Information Miner) es una plataforma de análisis de datos de código abierto que permite a los usuarios crear flujos de trabajo de ciencia de datos de manera visual e intuitiva. Se utiliza especialmente en campos como la investigación científica, análisis de big data, aprendizaje automático y procesamiento de grandes volúmenes de información. Su interfaz gráfica permite a científicos de datos, analistas e investigadores integrar diferentes fuentes de datos, realizar transformaciones complejas y desarrollar modelos predictivos sin necesidad de programación avanzada.

KNIME se destaca en sectores como la investigación biomédica, análisis financiero, marketing y predicción de riesgos. Su capacidad para integrar múltiples fuentes de datos, realizar análisis estadísticos complejos y crear modelos predictivos lo convierte en una herramienta versátil para profesionales que necesitan extraer conocimientos significativos de conjuntos de datos grandes y complejos.

6.1. Diagrama Bases de Datos No-SQL

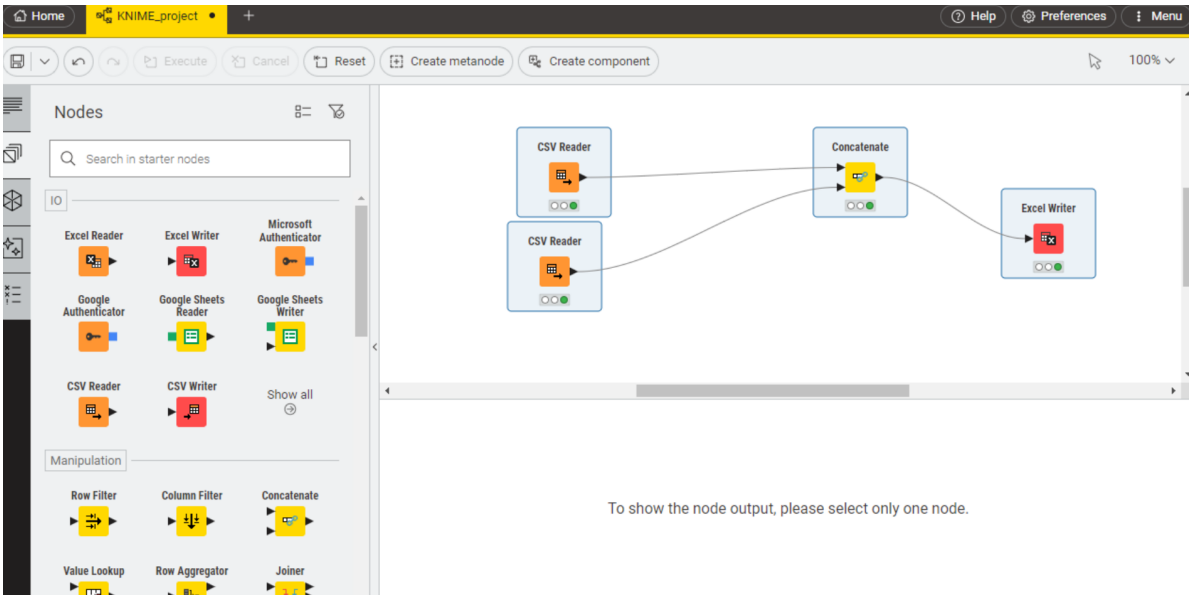

Figura 9: Base NoSQL de Saber Pro 2021

6.2. Ejemplo de aplicación de ETL y Bodega de Datos

En la sección 6.2 se muestra la evidencia de un ETL con el uso de KNIME para bases de datos NO SQL, en la imagen se puede apreciar la subida de un archivo tipo csv con 199 registros de nuestra base de datos de Saber Pro del año 2019 y cargamos otro archivo del mismo tipo con la información del año 2020, al cargarla nos saco un error al encontrar que el archivo tenia muchos datos y tuvo conflictos para leerlos, por lo cual, decidimos disminuir la cantidad de datos y columnas para poder realizar el ejercicio.

Finalmente cuando nos leyo los archivos lo que se quizo buscar es concatenar las bases de 2019 y 2020 con el fin de manejar o unificar los datos y poder manejarlos de manera mas eficiente, sin embargo, al intentar realizar el ejercicio con una base de datos medianamnete grande el software no logro leer ni concatenar los datos por lo cual se disminuyo la cantidad de datos cargados y de esta manera el flujo realizó el trabajo satisfactoriamente.

6.3. Automatización de Datos

La automatización de los datos a través de el software KNIME (Ordenes y Silipo,), se refleja la necesidad de automatizar procesos que necesitan ejecutarse repetidamente, como la limpieza, transformación y análisis de datos en múltiples conjuntos de datos o escenarios.

Es por ello que la automatización de datos en **KNIME** implica crear flujos de trabajo (*workflows*) que procesen datos de manera automática con pasos definidos. A continuación, se presenta un ejemplo utilizando uno de los archivos `SaberPro_Genericas_2021 MONGO.csv`.

1. Importar los datos

Para cargar los datos en KNIME:

- Usa el nodo `File Reader` o `CSV Reader`.
- Configura el nodo para cargar el archivo (`SaberPro_Genericas_2021 MONGO.csv`).
- Asegúrate de especificar correctamente el delimitador (por ejemplo, coma `,` o punto y coma `;`).

2. Exploración y limpieza de datos

Usa los siguientes nodos para explorar y limpiar los datos:

- `Data Explorer`: Analiza la estructura y distribuciones de los datos.
- `Missing Value`: Para tratar valores faltantes.
- `Column Filter`: Para eliminar columnas irrelevantes.
- `String Manipulation`: Para limpiar texto o transformar datos categóricos.

3. Automatización de transformaciones

Automatiza las transformaciones con nodos como:

- `Row Filter`: Filtra filas según condiciones específicas.
- `Column Expressions`: Realiza cálculos o transformaciones en columnas.
- `GroupBy`: Resume los datos (sumas, promedios, conteos, etc.).

4. Exportación de resultados

Para guardar los resultados procesados:

- Usa nodos como `CSV Writer` o `Excel Writer`.
- Configura la ruta de exportación para generar archivos de salida automáticamente.

5. Repetición con bucles (*Loops*)

Automatiza procesos repetitivos con nodos como:

- `Table Row to Variable Loop Start`: Itera sobre cada fila como variable.
- `Loop End`: Consolida los resultados tras finalizar el ciclo.

6. Programación del flujo

Para programar la ejecución automática:

- Usa nodos como `Timer Info` o `Wait`.
- Integra sistemas externos (por ejemplo, Python o APIs) si necesitas descargar o procesar datos dinámicamente.

7. Ejemplo del flujo

1. Leer el archivo `SaberPro_Genericas_2021 MONGO.csv`.
2. Limpiar las columnas irrelevantes.
3. Realizar cálculos como promedios o análisis por categorías.
4. Exportar el resultado final a un archivo CSV.

6.4. Integración de Datos (*Tercera entrega*)

Debido a la cantidad de datos que se manipulan en el proyecto, es necesario utilizar una plataforma manejasdora de bases de datos que tenga una capacidad de memoria grande y una escalabilidad apropiada para seguir alimentando el modelo una vezse cuenten con más resultados en años posteriores.

Al tener basesde datos separadas por año en el caso de las pruebas saber pro y tambien en las bases del rendimiento de colegios con las pruebas Saber 11, la integración de datos que se requiere necesita de una capacidad grande para manejar una cantidad masiva de datos, ya que como se quiere configurar la base de datos tendriamos 5 tablas con información relevante del estudiante, del colegio, de las pruebas, de la institución educativa de educación superior y de la familia del estudiante, en este sentido, es imperativo integrar los datos en una o dos bases de datos consolidadas que al menos nos unifique los años de estudio ya que como estan en este momento separadas por año resulta muy complejo realizar los análisis.

De acuerdo con la configuración de la base de datos y una vez se tenga la base consolidada en una gran base de datos podemos evidenciar la presencia de más de un millon de datos, esto se logro despues de realizar una unión en las bases de datos de Saber pro con la de rendimientos de colegios en Saber 11 relacionando el codigo dane del colegio, que es un codigo único de cada colegio del país y fue nuestra variable pivote para poder relacionar las dos bases.

Se hace necesario, no solo tener unas disparadores que organicen y clasifiquen los datos cuando se vayan cargando masivamente, si no tambien un software especializado para big data que me permita alojar grnades cantidades de datos de manera ordenada y de fail consulta y acceso, para esto utilizaremos triggers, vistas, procedimientos almacenados y funciones que nos van a permitir cumplir ocn estos requerimientos de integración de datos.

Así mismo, al hacer la union de las bases se pudo evidenciar que muchas de las variables de caracterización de los estudiantes que presentaron la prueba estan vacias a lo cual se pueden tomar la determinación de imputar estos datos o buscar completar esta información con ayuda de entidades especializadas en caracterizar la población como el DANE u otro organismo gubernamental.

7. Proximos pasos

A continuación se relacionan algunas consideraciones a tener encuenta:

- Consulta de otras bases de datos desagregadas y actualizadas ante el ICFES y el DANE, es decir, solicitar las bases de datos de Saber 11 y Pro anonimizadas con todas las variables, diccionarios, y sus respectivas llaves para poder optimizar los filtros y asegurar la calidad de la información.
- Revisión de los tipos de variables y sus características en el desarrollo socio-cultural colombiano para realizar análisis descriptivos para la comunidad en general.
- Dessarrollo y aplicación de un modelo de machine learning que permita predecir los resultados de las pruebas Saber Pro a partir de los resultados obtenidos por los estudiantes o instituciones en las Pruebas Saber 11.

8. Lecciones aprendidas

Finalmente en este apartado se presentarán las conclusiones que permitieron llevar a el desarrollo del documento:

- Se desarrollaron actividades de búsqueda y recopilación de los datos a traves de la página del ICFES y la Universidad Externado de Colombia cumpliendo con la politica de protección de datos anonimizados.
- Se aplicaron métodos de estádistica descriptiva usando el lenguaje de programación Python(González Duque,) a través de la plataforma de Google Colaborate (Dekeyser y Watson,) esto debido al volumen de los datos, en conjunto se realizarón pruebas a partir de muestras en las plataformas SQLDeveloper, Miro, MongoDB y KNIME.
- Desarrollando el trabajo en diferentes plataformas y momentos se determinó q ue se debe optar por tener licencias para garantizar la seguridad de la información y que permita el cargue de las bases que en conjunto suman más de $1'200,000$ aclarando que falta la solicitud a las instituciones mencionadas en el aprtado anterior (ICFES, DANE).

Referencias

Bock, D. B., Yager, S. E. (2005). Using the data modeling worksheet to improve novice data modeler performance. *Journal of Information Systems Education*, *16*(3), 341–350.

Dekeyser, S., Watson, R. (2006). Extending google docs to collaborate on research papers. *Toowoomba, Queensland, AU: The University of Southern Queensland, Australia*, *23*, 2008.

García, J., Martínez, A. (2018). *Contextos socioeconómicos y rendimiento académico en colombia: Un análisis comparativo*. Editorial Educativa.

González Duque, R. (2011). *Python para todos*. Creative Commons Reconocimiento.

Gupta, S. K. (2012). *Oracle advanced pl/sql developer professional guide*. Packt Publishing Ltd.

Leavitt, N. (2010). Will nosql databases live up to their promise? *Computer*, *43*(2), 12-14. doi: 10.1109/MC.2010.58

López, F., Gómez, C. (2021). Evaluación del desempeño académico en colombia: Un enfoque basado en modelos mixtos. *Journal of Educational Data Analytics*, *15*(2), 89–104.

Ordenes, F. V., Silipo, R. (2021). Machine learning for marketing on the knime hub: The development of a live repository for marketing applications. *Journal of Business Research*, *137*, 393–410.

Petrino, R. (2003). Artículo 11. protección de la honra y de la dignidad. *E. Alonso Regueira, La Convención Americana de derechos humanos y su proyección en el derecho argentino*, 203–217.

Rodríguez, M. (2019). *Modelos estadísticos avanzados para el análisis de datos educativos*. Publicaciones Académicas.

Ronderos, M. F. C. (2014). Legislación informática y protección de datos en colombia, comparada con otros países. *Inventum*, *9*(17), 32–37.

Strauch, C. (2011). Nosql databases. *Lecture Notes*. Descargado de https://examplesite.com

Vargas, J., Santos, M. (2018). Impacto de las variables contextuales en el rendimiento en las pruebas saber 11. *Revista Colombiana de Educación*, *27*(1), 67–82.