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]

18 de octubre 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 (?, ?, ?). 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(?, ?, ?).

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 (?, ?).

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** (?, ?) 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 del SMBD (Sistema Manejador de Bases de Datos)

Las entidades que se seleccionarón para el presente proyecto son las siguientes:

* Caracterización: se consideran variables que presentan la caracterización primaria del estudiante desde variables como fecha de nacimiento, género, departamento de residencia entre otras.

* Colegio: se consideran variables como código del dane del colegio permitiendo la caracterización de la institución con la que se inscribio el estudiante para presentar la prueba Saber 11.

* Hogar: Es la caracterización del hogar del estudiante y presenta variables tales como si tiene acceso a internet, educación de los padres, si tiene computador, entre otras.

* Examen: Para esta entidad se presenta la caracterización de la prueba Saber-Pro de cada estudiante considerando variables como departamento de presentación, municipio de presentación.

*IES: Corresponde a la caracterización de la universidad con la que el estudiante presento el examen Saber-Pro. *Financiación: Presenta la forma como el estudiante se financió los estudios de educación superior con variables como si es estudiante becado o con crédito, recursos propios, entre otras.

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 (?, ?).

* 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 (?, ?).

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.

Figura 1: Sistema de variables

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

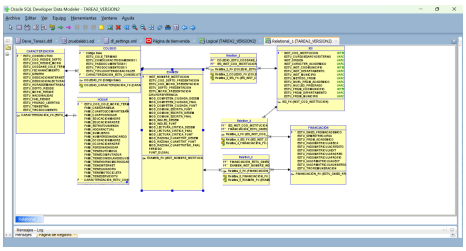

Figura 2: Sistema DDL base de datos

4.5. Código SQL - Manipulación de datos (DML)

```
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
);

```

```
59 ALTER TABLE examen ADD CONSTRAINT examen_pk PRIMARY KEY (
       inst_nombre_institucion);
60
61 CREATE TABLE financiaci n (
62     estu_snies_prgmacademico        INTEGER NOT NULL,
63     estu_semestrecursa              INTEGER,
64     estu_prgm_academico             VARCHAR2,
65     estu_pagomatriculabeca          CHAR(1),
66     estu_pagomatriculacredito       CHAR(1),
67     estu_pagomatriculaext           CHAR(1),
68     estu_pagomatriculapadres        CHAR(1),
69     estu_pagomatriculapropio        CHAR(1),
70     estu_valormatriculauext         INTEGER,
71     estu_valormatriculauniversidad INTEGER,
72     estu_tiporemuneracion           VARCHAR2(200 CHAR)
73 );
74
75 ALTER TABLE financiaci n ADD CONSTRAINT financiaci n_pk
       PRIMARY KEY (estu_snies_prgmacademico);
76
77 CREATE TABLE hogar (
78     estu_cod_cole_mcpio_termino       INTEGER NOT NULL,
79     fami_cabezafamilia                CHAR(1),
80     fami_cuantoscomparteba o           INTEGER,
81     fami_cuartoshogar                 INTEGER,
82     fami_educacionmadre               VARCHAR2(150 CHAR),
83     fami_educacionpadre               VARCHAR2(150 CHAR),
84     fami_estratovivienda              INTEGER,
85     fami_hogaractual                  VARCHAR2(150 CHAR),
86     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

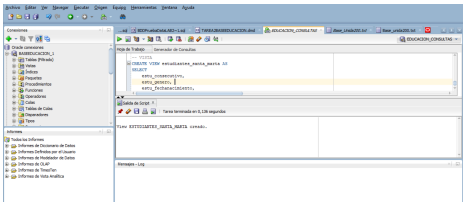

Figura 3: Vista de la base

```
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

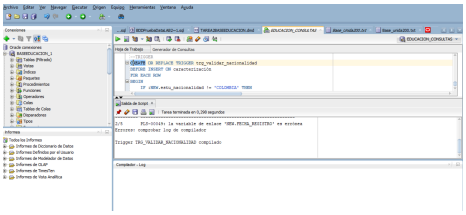

Figura 4: Trigger

–TRIGGER article listings xcolor

```
1 CREATE OR REPLACE TRIGGER trg_validar_nacionalidad
2
3 BEFORE INSERT ON caracterizaci n
4 FOR EACH ROW
5 BEGIN
6     IF :NEW.estu_nacionalidad != 'COLOMBIA' THEN
7         RAISE_APPLICATION_ERROR(-20001, 'Solo se permiten
              estudiantes de nacionalidad COLOMBIANA');
8     END IF;
9 END;
```