Pontificia Universidad Javeriana
Departamento de Ingeniería de Sistemas

Documento Proyecto
Visualización de datos
Análisis del sector educativo en Colombia

Nombre de integrantes:

Santiago Camilo Rey Benavides
William Andres Gomez Roa
Daniel Alfredo Vidal de Leon

Septiembre 22, 2024
Bogotá, Colombia

Tabla de contenido

Descripción del Sector Seleccionado

El sector educativo en Colombia, supervisado por el Ministerio de Educación Nacional (MEN) y las Secretarías de Educación Departamentales, enfrenta retos en la gestión y análisis de datos académicos. El MEN es responsable de formular políticas educativas y coordinar su implementación a nivel nacional, mientras que las Secretarías de Educación gestionan la educación a nivel regional. La implementación de un sistema de visualización de datos puede proporcionar *insights* cruciales para identificar tendencias, mejorar el rendimiento académico, optimizar la asignación de recursos y apoyar la formulación de políticas efectivas.

Objetivos de las Entidades

- **Promoción de la Calidad Educativa:** Implementar programas y recursos para mejorar la calidad de la educación, incluyendo capacitación para docentes y actualización de los currículos.

- **Fomento de la Equidad y la Inclusión:** Promover políticas que garanticen la equidad en el acceso y la calidad de la educación, atendiendo a poblaciones vulnerables y regiones con menor cobertura educativa.
- **Gestión de Recursos Educativos:** Administrar y distribuir recursos financieros y materiales para apoyar el desarrollo y la mejora del sistema educativo.
- **Formulación y Coordinación de Políticas Educativas:** Desarrollar y coordinar políticas y estrategias que mejoren la calidad y la equidad del sistema educativo a nivel nacional.
- **Garantizar el Acceso a la Educación:** Asegurar que todos los colombianos, sin importar su ubicación geográfica o condición socioeconómica, tengan acceso a una educación de calidad.
- **Monitoreo y Evaluación del Desempeño Educativo:** Evaluar y monitorear el rendimiento de los estudiantes y las instituciones educativas a través de pruebas estandarizadas y otros mecanismos de evaluación.

Contexto de Negocio

EduAnalytics Pro es una empresa innovadora dedicada a mejorar la calidad y equidad de la educación en Colombia mediante el análisis avanzado de datos educativos. Utilizando información de pruebas estandarizadas como SABER y SABER PRO, así como datos proporcionados por el MEN, *EduAnalytics* Pro se enfoca en proporcionar *insights* detallados para apoyar la toma de decisiones en el ámbito educativo.

Objetivos de *EduAnalytics Pro*

- **Mejorar la Calidad Educativa:** Utilizar datos de pruebas SABER y SABER PRO para identificar áreas de debilidad y fortaleza en el sistema educativo, y proporcionar recomendaciones para mejorar el rendimiento académico a nivel nacional.
- **Optimizar la Gestión Institucional:** Analizar la relación entre las características de las instituciones educativas (como jornada, valor de pensión, tipo de colegio) y el desempeño académico de los estudiantes, para apoyar la toma de decisiones en la gestión y planificación educativa.
- **Fomentar la Equidad Educativa:** Analizar los datos para identificar disparidades en el acceso y el rendimiento educativo entre diferentes regiones y grupos socioeconómicos.

Algunas Preguntas Objetivo

- ¿Cómo varía el rendimiento académico de los estudiantes según sus antecedentes familiares y socioeconómicos?
- ¿Existen disparidades significativas en el rendimiento académico entre diferentes regiones del país?
- ¿Qué áreas temáticas (por ejemplo, matemáticas, ciencias, lenguaje) muestran consistentemente debilidades en los resultados de SABER y SABER PRO?
- ¿Cómo influyen las características institucionales (como tipo de jornada y valor de la pensión) en el rendimiento académico de los estudiantes?

Objetivos de Visualización

- **Visualizar el rendimiento académico cruzado con antecedentes personales y socioeconómicos.**
    - **Métricas Clave:**
        - Promedio de calificaciones por materia.
        - Rendimiento por quintil socioeconómico.
- **Identificar disparidades en el rendimiento académico entre regiones y grupos socioeconómicos.**
    - **Métricas Clave:**
        - Índice de desigualdad por región.
        - Promedio de calificaciones por región/grupo socioeconómico.
- **Identificar patrones de rendimiento en SABER y SABER PRO.**
    - **Métricas Clave:**
        - Número de estudiantes en niveles avanzados (SABER/SABER PRO).
        - Comparación de rendimiento a lo largo del tiempo.
- **Analizar la relación entre características institucionales y rendimiento académico.**
    - **Métricas Clave:**

        - Rendimiento académico por tipo de institución.
        - Correlación entre valor de la pensión y rendimiento académico.
- **Explorar cómo factores personales, como el nivel educativo, el género, etc., influyen en el rendimiento académico de los estudiantes.**
    - **Métricas Clave:**
        - Rendimiento académico promedio por nivel educativo.
        - Comparación del rendimiento entre géneros en diferentes materias.

Datasets

1. **Saber 11: 2020-2023**

Se unieron 8 conjuntos de datos de las pruebas Saber 11 realizadas desde el año 2020-1 hasta el año 2023-2 con el objetivo de recopilar información a lo largo de estos años. Los conjuntos de datos fueron limpiados y combinados en uno solo. Se creó un pipeline para la limpieza y unión de estos datos, permitiendo muestrear aleatoriamente 1,000 filas de cada conjunto. Sin embargo, este número puede aumentarse fácilmente cuando se realicen las visualizaciones, si es necesario. Después se limpia el conjunto de datos, eliminando las filas con muchos campos vacíos.

- **Columnas:** 77
- **Filas:** 7.432
- **Variables de interés:**
    - ESTU_GENERO
    - ESTU_DEPTO_RESIDE
    - ESTU_MCPIO_RESIDE
    - FAMI_EDUCACIONMADRE
    - FAMI_TIENEINTERNET
    - COLE_BILINGUE
    - PUNT_C_NATURALES
    - PUNT_INGLES
    - PUNT_MATEMATICAS

        - PUNT_GLOBAL

2. **Saber Pro: 2011-2022**

Se unieron 5 conjuntos de datos de las pruebas Saber Pro realizadas desde el año 2020 hasta el año 2023 con el objetivo de recopilar información a lo largo de estos años. Los conjuntos de datos fueron limpiados y combinados en uno solo. Se creó un pipeline para la limpieza y unión de estos datos, permitiendo muestrear aleatoriamente 1,000 filas de cada conjunto. Sin embargo, este número puede aumentarse fácilmente cuando se realicen las visualizaciones, si es necesario. Después se limpió el conjunto de datos, eliminando las filas con muchos campos vacíos.

- **Columnas:** 67
- **Filas:** 5.000
- **Variables de interés:**
    - ESTU_DEPTO_RESIDE
    - ESTU_VALORMATRICULAUNIVERSIDAD
    - FAMI_EDUCACIONPADRE
    - ESTU_PRGM_ACADEMICO
    - MOD_LECTURA_CRITICA_PUNT
    - MOD_RAZONA_CUANTITAT_PUNT
    - MOD_COMUNI_ESCRITA_PUNT
    - MOD_INGLES_PUNT
    - PUNT_GLOBAL
    - PERCENTIL_GLOBAL

3. **Estadísticas en Educación Preescolar, Básica y Media por Municipio: 2011-2022**

Se tomó el único conjunto de datos correspondiente a las Estadísticas de Educación del MEN, a nivel preescolar, básica y media por municipio entre los años 2011 y 2022. Para eliminar las filas con muchos campos vacíos se limpiaron estos datos. Asimismo, se muestrearon aleatoriamente 3,000 filas antes de la limpieza para trabajar con menos datos. Sin embargo, mediante el pipeline desarrollado, es posible aumentar fácilmente el número de filas.

- **Columnas:** 41
- **Filas:** 3.000
- **Variables de interés:**
  - POBLACIÓN_5_16
  - MUNICIPIO
  - DEPARTAMENTO
  - COBERTURA_NETA
  - APROBACIÓN_SECUNDARIA
  - REPITENCIA_SECUNDARIA
  - DESERCIÓN_SECUNDARIA

Relacionando los *Datasets*

- **FAMI_EDUCACIONPADRE y FAMI_EDUCACIONMADRE:** Estos datos podrían correlacionarse con el rendimiento académico, ya que se ha demostrado que la educación de los padres influye en el desempeño de los estudiantes.
- **Resultados de las pruebas de SABER PRO** (por ejemplo, PUNT_GLOBAL) pueden correlacionarse con indicadores de MEN, como DESERCIÓN y APROBACIÓN, para entender cómo la calidad educativa y las condiciones del entorno escolar impactan en el rendimiento.
- **Los puntajes promedio de Saber 11** en diferentes departamentos o municipios, correlacionándolo con los indicadores educativos del MEN para identificar patrones geográficos.

Visualizaciones Planeadas

- Gráficos de dispersión (calificaciones vs. quintil).
- Barras apiladas (rendimiento por características personales).
- Barras comparativas (rendimiento por región y quintil socioeconómico).
- Gráficos de línea (evolución del rendimiento).
- *Heatmaps* (rendimiento por región).

Conclusiones

- Los datos revelan disparidades significativas en el rendimiento académico entre las diferentes regiones de Colombia. Estas desigualdades subrayan la importancia de políticas educativas específicas para zonas con menor rendimiento académico, como aquellas con menores recursos económicos o con menos acceso a tecnología.
- Los antecedentes familiares y el nivel socioeconómico de los estudiantes tienen una clara correlación con su rendimiento académico. Las visualizaciones muestran cómo los estudiantes de familias con mayor nivel educativo y acceso a internet tienden a obtener mejores resultados en las pruebas SABER y SABER PRO. Esto refuerza la importancia de programas de apoyo para estudiantes de entornos más vulnerables.
- Las características de las instituciones educativas, como el tipo de jornada y el valor de la pensión, también influyen en los resultados académicos. Las instituciones con mayor inversión en recursos y jornadas extendidas tienden a mostrar un mejor rendimiento. Esto sugiere que la optimización de la gestión institucional y la mejora de las infraestructuras escolares pueden ser clave para incrementar el rendimiento académico en el país.
- El análisis de los datos reveló que existen disparidades notables en el rendimiento académico entre diferentes regiones y quintiles socioeconómicos. Los estudiantes de zonas rurales o de menores ingresos tienden a tener un rendimiento inferior en las pruebas SABER y SABER PRO en comparación con aquellos de zonas urbanas y con mayores recursos. Esto subraya la necesidad de políticas educativas más inclusivas que garanticen recursos y oportunidades equitativas, independientemente del contexto geográfico o socioeconómico.
- La educación de los padres influye de manera importante en el rendimiento académico de los estudiantes: Los datos sugieren una fuerte correlación entre el nivel educativo de los padres y el rendimiento académico de los estudiantes. Aquellos cuyos padres tienen niveles más altos de educación tienden a obtener mejores resultados en las pruebas SABER, lo que resalta la importancia de fomentar un entorno educativo de apoyo desde el hogar. Esto sugiere que las estrategias educativas no deben limitarse solo a los estudiantes, sino también involucrar a las familias para maximizar el impacto en el desempeño académico.

Lecciones aprendidas

- La integración de múltiples fuentes de datos mejora el análisis educativo: Al combinar datos de diferentes fuentes como SABER, SABER PRO y estadísticas municipales, se obtiene una visión más completa del sistema educativo en Colombia. La correlación de variables como el nivel educativo de los padres, el rendimiento académico y los indicadores regionales permite identificar patrones que no serían evidentes con datasets aislados. Esto demuestra que el análisis de datos multidimensional es crucial para tomar decisiones informadas en el sector educativo.
- La precisión y completitud de los datos impactan directamente en la calidad de los insights generados por el modelo analítico. En el contexto educativo, trabajar con datos incompletos o mal estructurados, como los de encuestas socioeconómicas o registros institucionales, puede llevar a conclusiones erróneas o decisiones equivocadas. Para obtener resultados útiles, es esencial realizar una limpieza y validación rigurosa de los datos, asegurándose de que las variables clave estén bien definidas y alineadas con los objetivos del análisis. Esto asegura que los hallazgos reflejen la realidad del sistema educativo y sirvan como base sólida para la toma de decisiones.

Referencias

- https://www.icfes.gov.co/data-icfes/
- www.mineducacion.gov.co
- www.dane.gov.co
- https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/nudc-7mev/about_data

Anexos

- **join-df-SBPRO.py**: Archivo que abre 5 conjuntos de datos relacionados con las pruebas Saber Pro, los une y los limpia en un nuevo archivo muestreando 1.000 filas aleatorias de cada conjunto de datos original.
- **joint_datasets-SB11.py**: Archivo que abre 8 conjuntos de datos de las pruebas Saber 11, los junta en uno solo y realiza una limpieza de las filas con gran cantidad

de campos vacíos. Finalmente muestrea 1.000 filas de cada conjunto de datos original.

- **open-clean-MEN.py**: Abre el conjunto de datos, elimina las filas con gran cantidad de campos vacíos según una inspección visual, y finalmente muestre 3.000 filas del conjunto de datos original.
- **MEN_EDUCACION_limpio.csv:** Conjunto de datos limpio y pequeño para pruebas de visualización.
- **SB11-clean.txt:** Conjunto de datos limpio y pequeño para pruebas de visualización.
- **SBPro-clean.txt:** Conjunto de datos limpio y pequeño para pruebas de visualización.