Informe: Estrategias educativas para la población colombiana.

1. Resumen Ejecutivo:

- Objetivo del proyecto: Analizar los factores que influyen en el desempeño de los estudiantes en las pruebas Saber 11, con el fin de identificar oportunidades de mejora y optimizar los resultados.
- Problema o necesidad abordada: Bajo rendimiento en las pruebas Saber 11 en ciertos contextos educativos, lo que limita las oportunidades académicas y profesionales de los estudiantes.
- Hallazgos clave:
    - Los estudiantes de colegios con jornadas completas y acceso a recursos como computador, internet y lavadora en sus hogares tienden a obtener mejores resultados.
    - Los colegios rurales presentan mayores desafíos, evidenciando la necesidad de fortalecer las condiciones educativas en estas zonas.
- Impacto esperado: Identificar áreas críticas de intervención para mejorar las condiciones educativas y, en consecuencia, elevar los resultados de las pruebas Saber 11 en los grupos más vulnerables.

2. Contexto del Proyecto:

- Pregunta de investigación:
    - ¿Cuáles son los factores asociados a los buenos y malos resultados en las pruebas Saber 11?
- Metodología:
    - Fuente de datos: Se utilizó un conjunto de datos de las pruebas Saber 11 desde el año 2010 hasta el 2022, que incluye información detallada sobre el colegio, la familia, el estudiante y los puntajes obtenidos.
    - Análisis de agrupamiento: Mediante el algoritmo de agrupación K-Means, se segmentaron los datos en grupos homogéneos basados en características de los colegios, las familias y los estudiantes.
    - Evaluación de resultados: Se analizó el desempeño en las pruebas Saber 11 de cada grupo, identificando aquellos con los mejores y peores resultados.
    - Identificación de características clave: Se determinaron las características distintivas de los grupos con alto y bajo rendimiento.
    - Recomendaciones: A partir de los hallazgos, se identificaron los aspectos que requieren mejora y aquellos que contribuyen a un mejor desempeño en las pruebas.

3. Descripción de los Datos:

- Fuentes de datos:
  Los datos fueron obtenidos de un repositorio de datos abiertos de Colombia, proporcionando acceso a información detallada sobre las pruebas Saber 11.
- Características de los datos:

- El conjunto de datos está estructurado en formato tabular, con aproximadamente 7,11 millones de filas y 51 columnas.
- Las columnas incluyen información sobre:
  - El estudiante: Datos personales y demográficos.
  - La familia: Características socioeconómicas y educativas.
  - El colegio: Información institucional y contextual.
  - Resultados de las pruebas: Puntajes numéricos obtenidos en las pruebas Saber 11.

- Preprocesamiento:
  Para simplificar el análisis y la agrupación, se realizaron las siguientes transformaciones en los datos:
  1. Condensación de variables categóricas:
     - Columna FAMI_CUARTOSHOGAR: Los valores 'Seis', 'Diez o más', 'Ocho', 'Seis o más', 'Siete', 'Nueve' se consolidaron en la categoría 'Seis o más'.
     - Columna FAMI_PERSONASHOGAR:
       - 'Una' y 'Dos' se agruparon como '1 a 2'.
       - 'Tres' y 'Cuatro' se agruparon como '3 a 4'.
       - 'Cinco' y 'Seis' se agruparon como '5 a 6'.
       - 'Siete' y 'Ocho' se agruparon como '7 a 8'.
       - 'Nueve', 'Diez', 'Once' y 'Doce o más' se agruparon como '9 o más'.
     - Columnas FAMI_EDUCACIONMADRE y FAMI_EDUCACIONPADRE:
       - 'No sabe', 'No disponible', 'No Aplica' y 'Ninguno' se consolidaron como 'No disponible'.
       - 'Primaria incompleta' y 'Primaria completa' se agruparon como 'Primaria'.
       - 'Secundaria (Bachillerato) incompleta' y 'Secundaria (Bachillerato) completa' se agruparon como 'Secundaria'.
       - 'Técnica o tecnológica incompleta' y 'Técnica o tecnológica completa' se agruparon como 'Técnica'.
       - 'Educación profesional incompleta' y 'Educación profesional completa' se agruparon como 'Profesional'.
  2. Creación de nuevas variables:
     - Columna ESTU_EDAD:
       Se calculó restando el año de presentación de la prueba y el año de nacimiento del estudiante. Luego, se categorizó en:
       - '15 a 18',
       - '19 a 25',
       - '26 o más',
       - 'Otro'.
     - Columna ESTU_MOVI_EXAM:
       Se creó comparando los valores de ESTU_MCPIO_PRESENTACION y ESTU_MCPIO_RESIDE. Se asignó 'SI' si coinciden y 'NO' en caso contrario.

4. Análisis y Hallazgos:

- Técnicas utilizadas: Se identifican la cantidad de valores nulos de acuerdo al segmento de información (Familia, Estudiante, Colegio) obteniendo que:
    - La cantidad de datos completos de los colegios es 86.82%.
    - La cantidad de datos completos de las familias es 95.05%.
    - La cantidad de datos completos de los estudiantes es 100%.

  Con 7’109.704 registros, la descripción estadística de los colegios es:

Descripción	COLE_AREA_UBICACION	COLE_BILINGUE	COLE_CALENDARIO	COLE_CARACTER	COLE_GENERO	COLE_JORNADA	COLE_NATURALEZA	COLE_SEDE_PRINCIPAL
Valores únicos	3	3	4	5	4	7	3	3
Moda	URBANO	N	A	ACADÉMICO	MIXTO	MAÑANA	OFICIAL	S
Frecuencia moda / Cantidad	0,86	0,85	0,96	0,53	0,96	0,48	0,72	0,97

Para los colegios se identificaron 4 agrupaciones con las siguientes descripciones:

- Grupo 1: Cuenta con 3’773.304:

Descripción	COLE_AREA_UBICACION	COLE_BILINGUE	COLE_CALENDARIO	COLE_CARACTER	COLE_GENERO	COLE_JORNADA	COLE_NATURALEZA	COLE_SEDE_PRINCIPAL
Valores únicos	2	3	3	4	3	6	1	2
Moda	URBANO	N	A	TÉCNICO/ACADÉMICO	MIXTO	MAÑANA	OFICIAL	S
Frecuencia moda / Cantidad	1.00	0.91	1.00	0.52	0.97	0.53	1.00	0.97

- Grupo 2: Cuenta con 1’764.754:

Descripción	COLE_AREA_UBICACION	COLE_BILINGUE	COLE_CALENDARIO	COLE_CARACTER	COLE_GENERO	COLE_JORNADA	COLE_NATURALEZA	COLE_SEDE_PRINCIPAL
Valores únicos	2	3	4	5	4	7	2	3
Moda	URBANO	N	A	ACADÉMICO	MIXTO	COMPLETA	NO OFICIAL	S
Frecuencia moda / Cantidad	1.00	0.79	0.89	0.83	0.93	0.43	1.00	1.00

- Grupo 3: Cuenta con 1’000.937:

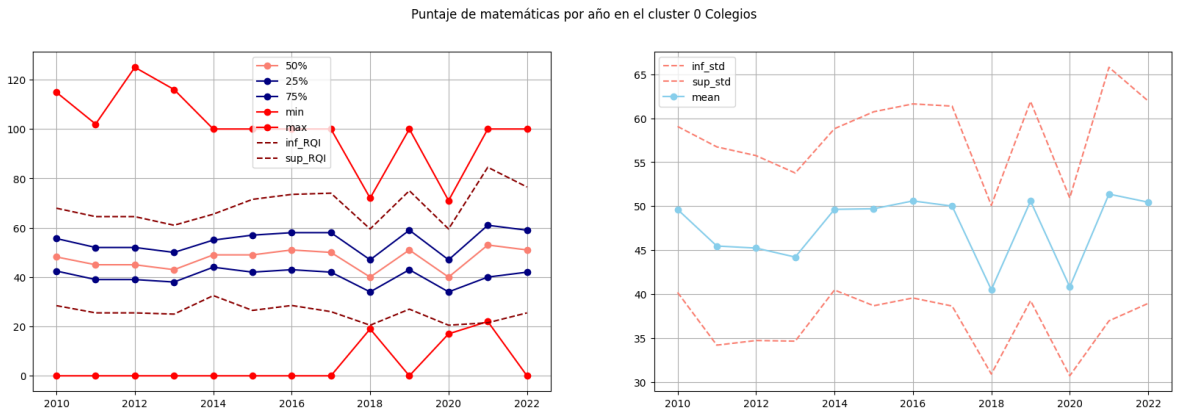
Descripción	COLE_AREA_UBICACION	COLE_BILINGUE	COLE_CALENDARIO	COLE_CARACTER	COLE_GENERO	COLE_JORNADA	COLE_NATURALEZA	COLE_SEDE_PRINCIPAL
Valores únicos	2	3	4	5	3	6	2	2
Moda	RURAL	N	A	ACADÉMICO	MIXTO	MAÑANA	OFICIAL	S
Frecuencia moda / Cantidad	1.00	0.73	0.97	0.54	0.99	0.56	0.90	0.94

- Grupo 4: Cuenta con 570.709:

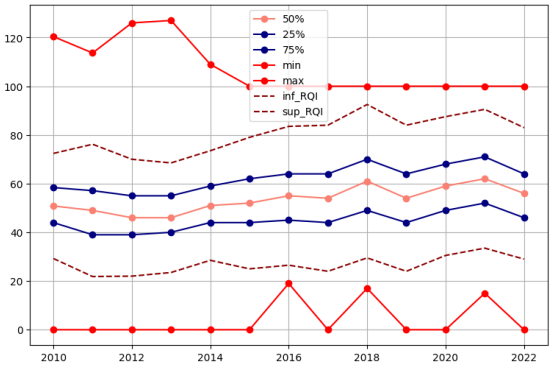
Descripción	COLE_AREA_UBICACION	COLE_BILINGUE	COLE_CALENDARIO	COLE_CARACTER	COLE_GENERO	COLE_JORNADA	COLE_NATURALEZA	COLE_SEDE_PRINCIPAL
Valores únicos	1	3	2	1	3	6	2	2
Moda	URBANO	N	A	TÉCNICO	MIXTO	MAÑANA	OFICIAL	S
Frecuencia moda / Cantidad	1.00	0.93	0.95	1.00	0.95	0.54	0.81	0.96

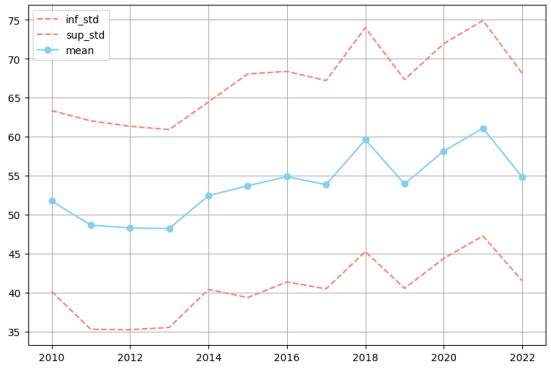
- Resultados clave:

  De acuerdo a las agrupaciones obtenidas de los colegios, se observa que el grupo 2 (Cluster 1) presenta mejores resultados en matemáticas, caracterizando este grupo, se observa, a diferencia de los otros, que presenta jornada continua y la naturaleza de los colegios es No Oficial.

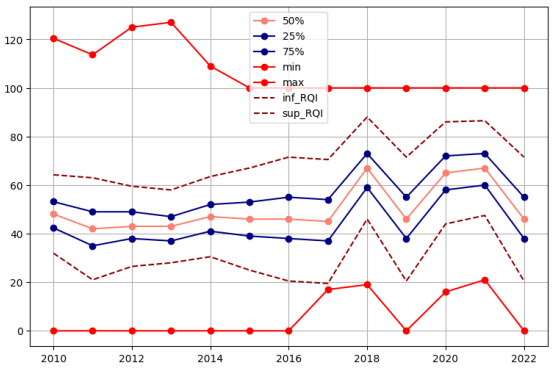


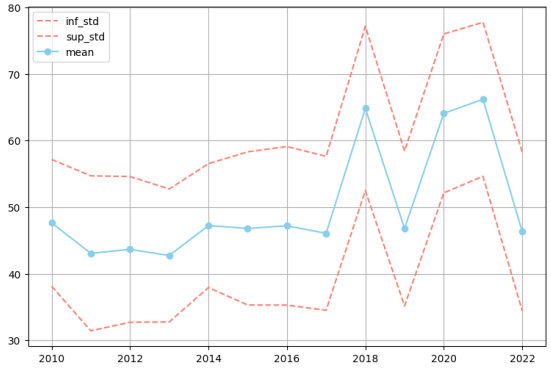
Puntaje de matemáticas por año en el cluster 1 Colegios



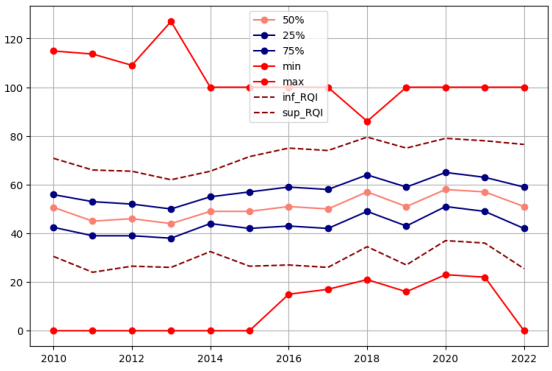


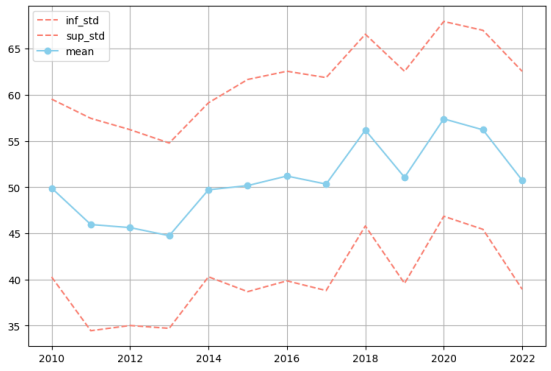
Puntaje de matemáticas por año en el cluster 2 Colegios





Puntaje de matemáticas por año en el cluster 3 Colegios





5. Conclusiones:

- Factores asociados a mejores resultados:
    - Los estudiantes de colegios de jornada completa y de carácter No Oficial tienden a obtener puntajes más altos en las pruebas Saber 11.
    - Las familias que cuentan con recursos como computadora, internet y lavadora muestran una correlación positiva con mejores resultados académicos.

    - La edad idónea para presentar el examen se encuentra entre los 15 y 18 años, grupo que demuestra un desempeño superior.
- Áreas de mejora:
    - Los colegios rurales presentan los puntajes más bajos, lo que evidencia la necesidad de fortalecer las condiciones educativas en estas zonas para reducir la brecha de rendimiento.