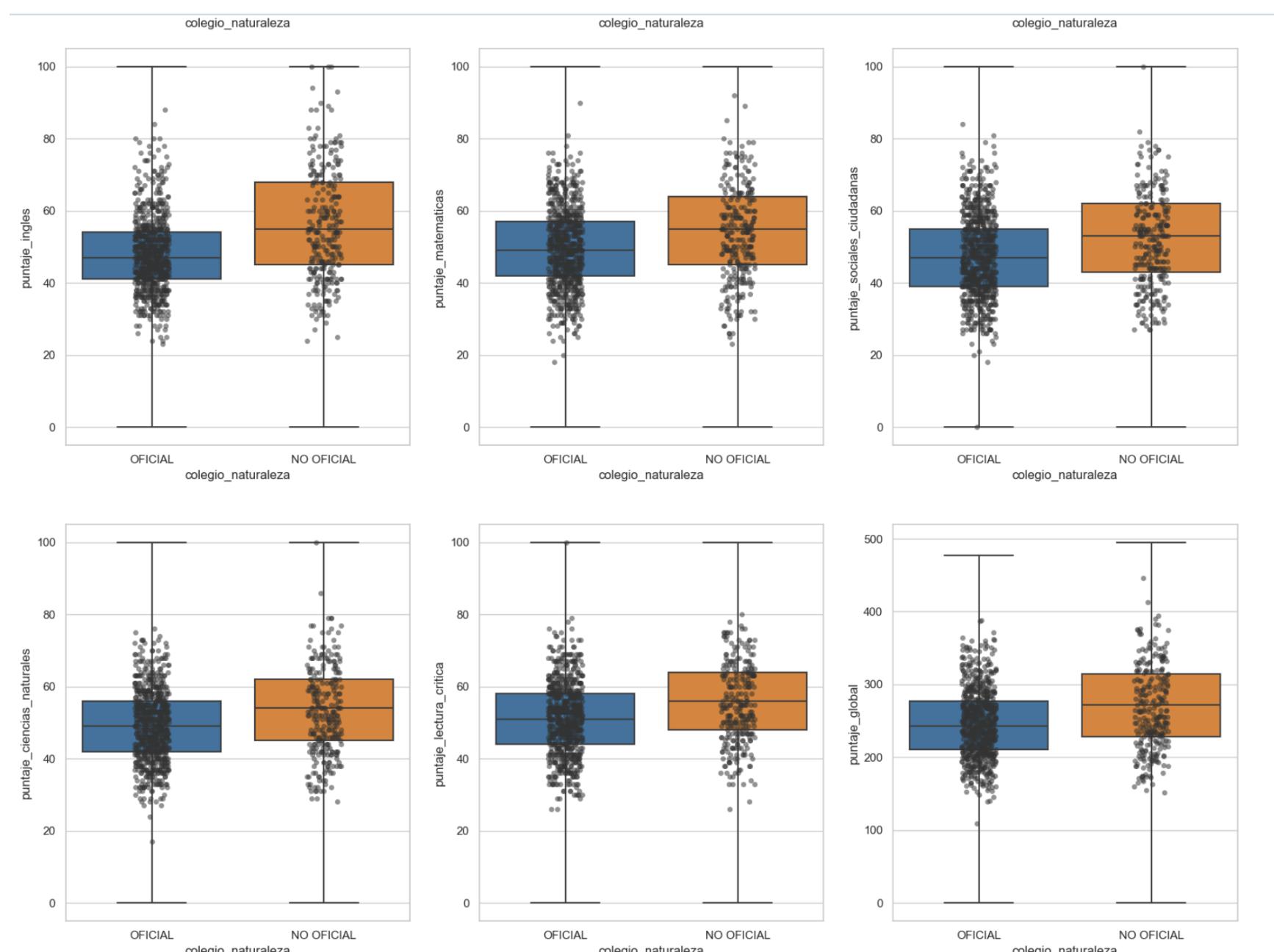


PredICFES: Visualizando el Futuro Académico en Colombia

Análisis de datos

Relación del puntaje total con factores demográficos

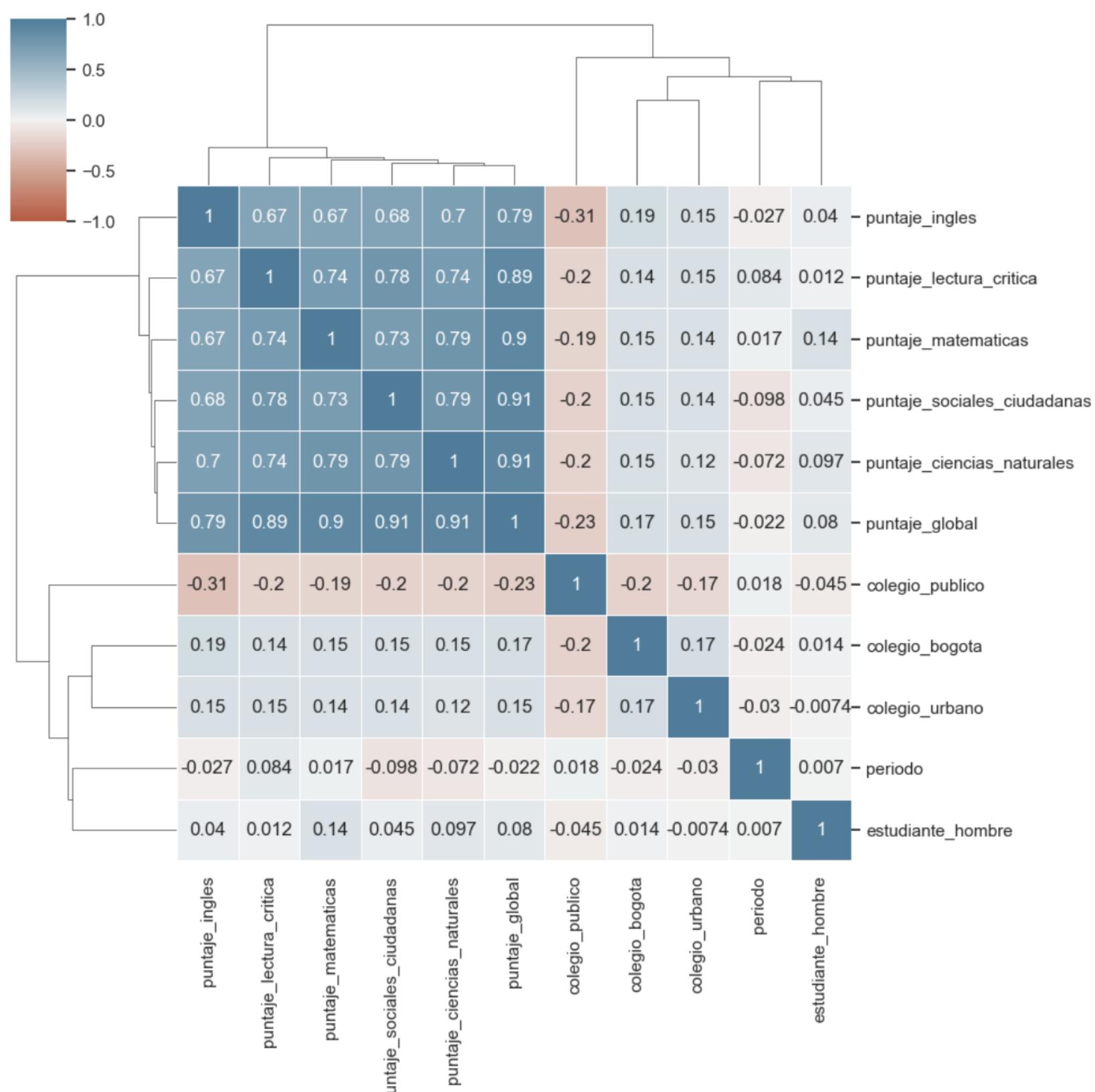


Al observar los boxplots de cada área, se evidencia una clara tendencia: los colegios privados superan consistentemente a los colegios oficiales en todos los ámbitos evaluados.

El hecho de que los boxplots de los colegios privados estén consistentemente ubicados por encima de los de los colegios oficiales indica que, en promedio, los estudiantes de los colegios privados obtienen puntajes más altos en todas las áreas evaluadas. Esto sugiere posibles diferencias en la calidad de la enseñanza, los recursos disponibles o el entorno educativo entre los dos tipos de instituciones.

A pesar de esta disparidad en los puntajes promedio, es importante tener en cuenta que la media en todas las áreas sigue siendo de 50, lo que indica que los exámenes están diseñados para ser equitativos en términos de dificultad y nivel de habilidad requerido. Sin embargo, la diferencia en los puntajes promedio entre colegios privados y oficiales puede plantear preguntas importantes sobre la equidad en el acceso a una educación de calidad y sobre cómo mejorar los resultados educativos en todas las instituciones.

Correlación entre variables calculadas



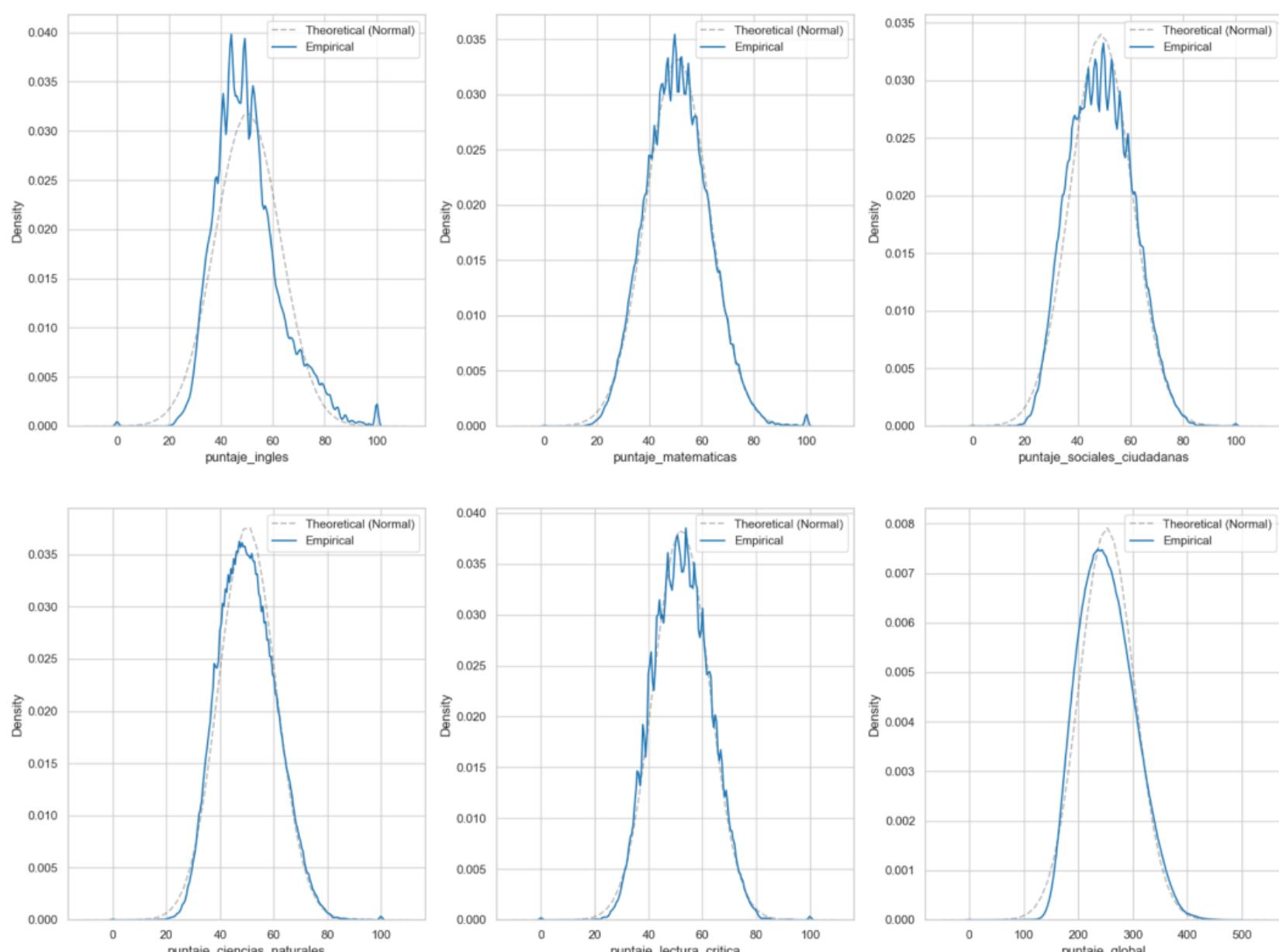
Se observa que, en general, los estudiantes que destacan en una asignatura tienden a hacerlo también en otras, lo que sugiere una comprensión integral de los conceptos educativos. Sin embargo, se ha identificado una discrepancia notable en el colegio público: el puntaje de inglés es bajo, con un coeficiente de correlación de -0.31. De forma general, aquellos estudiantes que provienen de colegios públicos, por fuera de Bogotá y/o estudian en entornos no urbanos, suelen tener mayores dificultades con la prueba. Es interesante notar a su vez que estas características están correlacionadas a su vez entre sí.

Además, se ha encontrado que el género no influye significativamente en el rendimiento académico, excepto en matemáticas. En esta área, se observa una influencia considerable del género, con un desempeño generalmente más alto entre los estudiantes masculinos en comparación con las estudiantes femeninas. Esta diferencia plantea

preguntas importantes sobre los métodos de enseñanza y los recursos disponibles para abordar las disparidades de género en el rendimiento académico en matemáticas.

También se ha notado que el periodo influye ligeramente en el rendimiento en sociales y ciencias naturales. Es alentador ver que, fuera de la diferencia de género en matemáticas, el desempeño en otras áreas es más equitativo, reflejando un entorno educativo que promueve la igualdad de oportunidades para todos los estudiantes.

Distribución de los puntajes de los estudiantes



En el análisis de los resultados académicos en áreas como inglés, matemáticas, competencias ciudadanas, ciencias naturales y lectura crítica, surge una observación destacada: la tendencia de los puntajes a ajustarse a una distribución normal, con una media común de 50. Este fenómeno, que se repite en múltiples áreas de evaluación, proporciona una perspectiva valiosa sobre la dinámica y el rendimiento del sistema educativo.

La consistencia en los puntajes, alineada con una distribución normal, sugiere que el sistema de evaluación utilizado es sólido y equitativo, capturando de manera efectiva las habilidades de los estudiantes en diversos campos del conocimiento. Sin embargo, más allá de la uniformidad en la media, la distribución normal revela la diversidad de habilidades presentes entre los estudiantes. Algunos destacan significativamente por encima de la media en ciertas áreas, mientras que otros pueden encontrarse por debajo, reflejando la variedad de aptitudes y niveles de competencia presentes en la población estudiantil.

La distribución normal de los puntajes en diversas áreas sugiere un equilibrio en el programa educativo, que ofrece una variedad de asignaturas y actividades para el desarrollo integral de los estudiantes. Además, sirve como punto de referencia para evaluar la efectividad de las intervenciones educativas y orientar la mejora continua del sistema educativo en su conjunto.

Ingeniería de Datos

De forma general, empleamos diversas tácticas para poder manejar el volumen de información suministrado por el ICFES (más de 7 millones de registros de resultados). A continuación describiremos algunas de ellas:

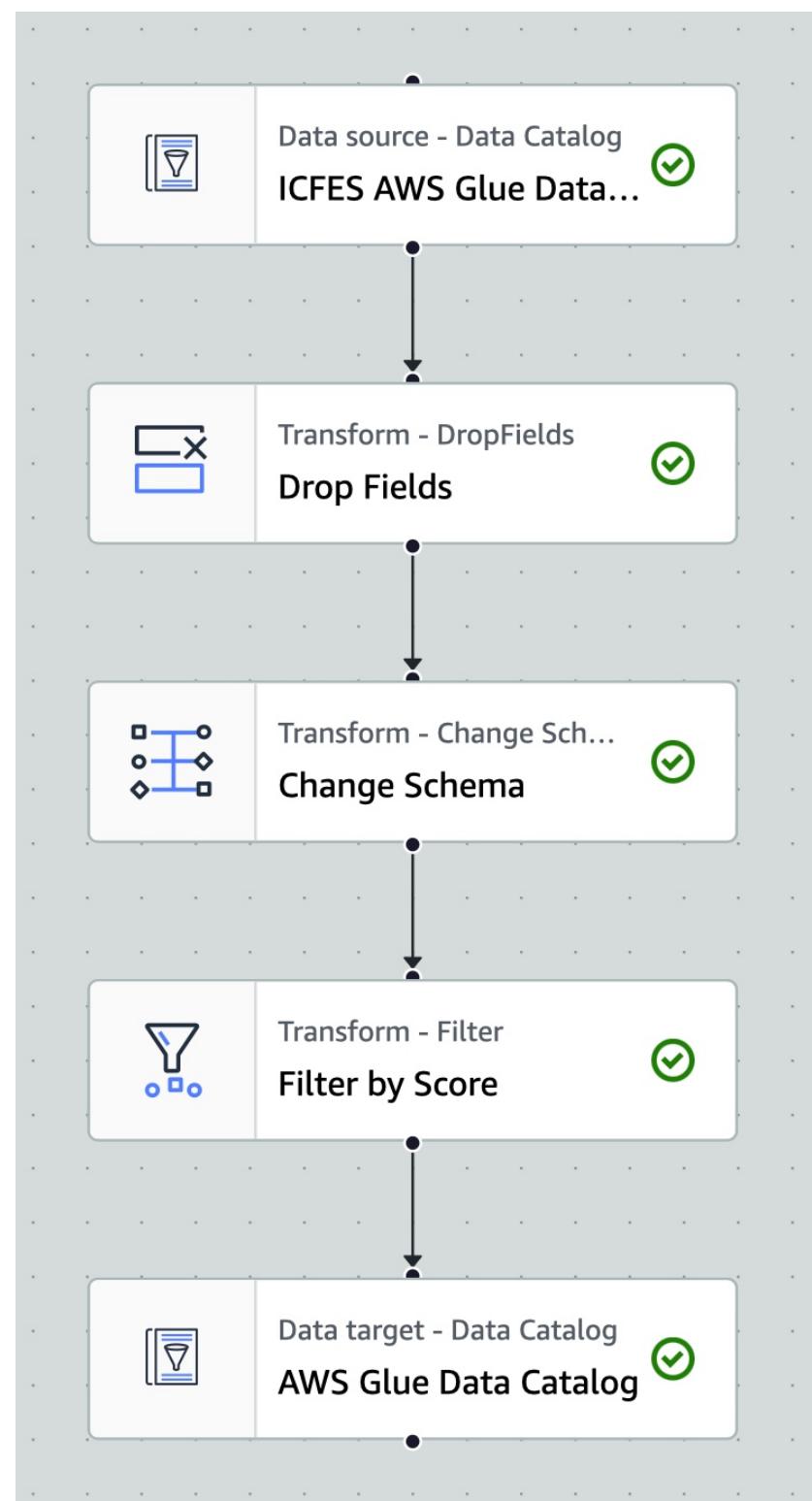
Descarga de los archivos

Programamos un script automático que descargara los datos desde la API proporcionada por el ICFES, teniendo en cuenta el límite de datos que se pueden pedir en una única solicitud y el cálculo del offset entre solicitudes. Luego, usamos Crawlers de Glue para manipular todos los archivos de CSV generados (100+) como si fueran una única base de datos.

Extracción, Transformación y Carga

Utilizamos las herramientas de ETL de Glue como una primera aproximación a los datos, debido a que el volumen masivo de los mismos impide realizar operaciones complejas sobre los mismos de forma local. Por medio de Glue, realizamos los siguientes pasos:

1. Leímos los datos organizados por el Crawler desde el bucket S3 de archivos CSV descargados del ICFES.
2. Solucionamos errores de columnas con tipos de datos indefinidos al importar.
3. Eliminamos campos que no requerimos en nuestro análisis, por ejemplo, la información de la encuesta demográfica.
4. Re-estructuramos el esquema utilizando nombres de columnas más claros.
5. Filtramos los valores válidos de las columnas más importantes de nuestro análisis, a saber, los puntajes. En este proceso, también eliminamos nulos.
6. Guardamos los datos nuevamente en una nueva base de datos de Glue, desde la cual hicimos consultas y exportamos.



Creación de vistas personalizadas

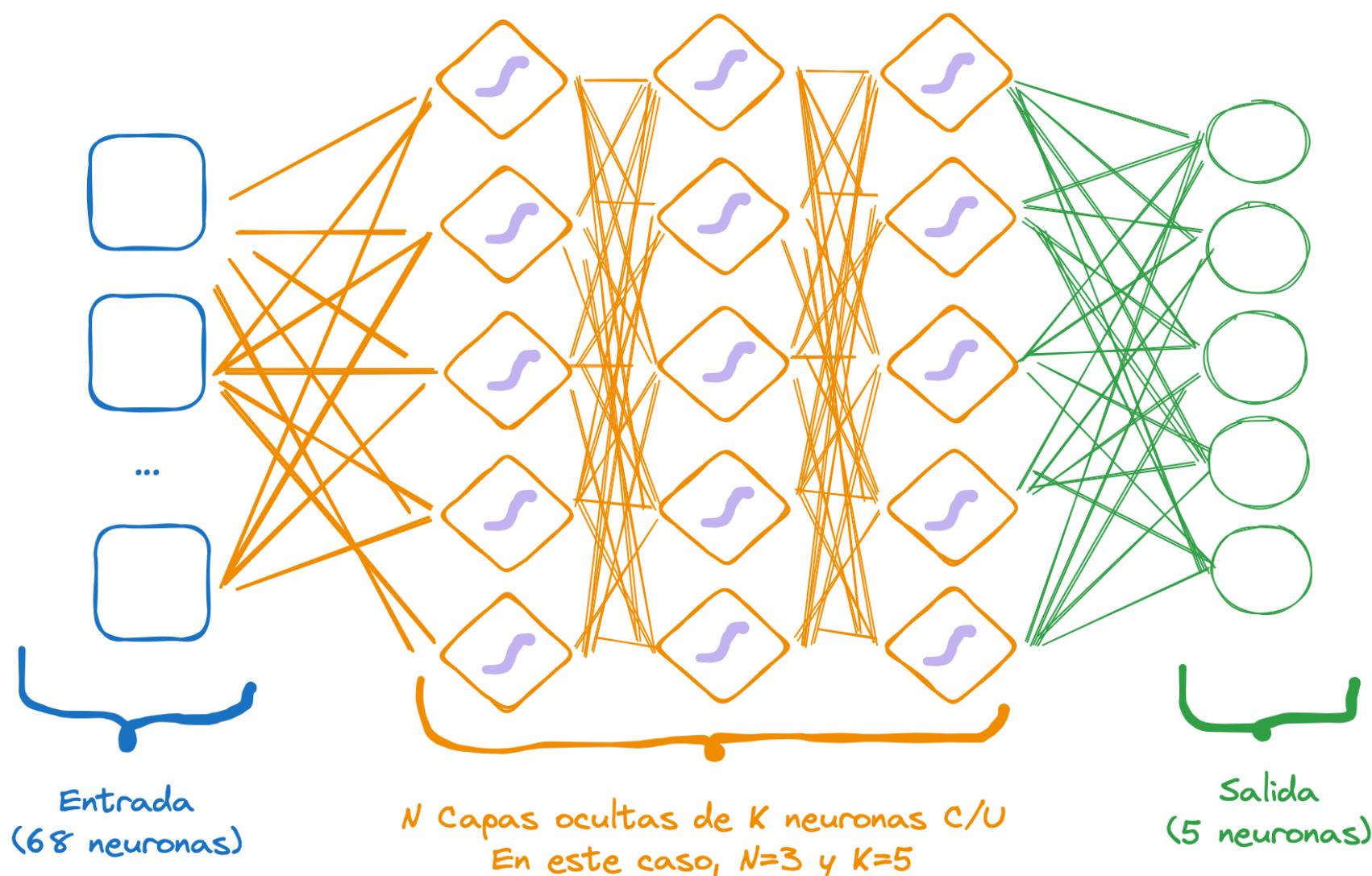
Consultar nuestra base de datos completa es un proceso tedioso que se puede tardar hasta 20 minutos únicamente en la operación de lectura. Esto se debe a la masiva cantidad de información que almacena la base de datos. Con el motivo de agilizar este proceso y ofrecer una mejor experiencia de usuario, creamos vistas materializadas sobre la base de datos completa, que permiten acceder rápidamente a un subconjunto de los datos muy particular. Creamos estas vistas para las consultas más comunes desde el front-end de nuestra aplicación.

Modelos de aprendizaje automático

Se empleó un enfoque basado en redes neuronales para predecir el desempeño en las distintas competencias del examen ICFES Saber 11 (Matemáticas, Lectura Crítica, Ciencias Naturales, Ciencias Sociales e Inglés). El desempeño global de la prueba se puede calcular a partir del desempeño en cada competencia, por lo que seguir este enfoque puede proporcionarnos información más precisa para el entrenamiento de los modelos. Por su lado, la naturaleza de estos datos es la de un porcentaje, lo que la hace apta para ser modelada como una función de activación final *sigmoide*.

Siguiendo esta idea, se planteó una red neuronal sencilla compuesta por varias capas densas. Inicialmente, se reciben las variables preprocesadas y se pasan por una capa con 512 neuronas. Luego, se sigue procesando la información por otras n capas con la misma cantidad de neuronas k, hasta llegar a la capa de salida con 5 neuronas y función de activación sigmoide.

Determinamos como hiperparámetros relevantes la cantidad de capas, el número de neuronas por capa y la función de activación de las capas intermedias. Evaluamos cada uno de estos parámetros de forma independiente con el fin de determinar su desempeño. Utilizamos el framework MLFlow para documentar los resultados de cada experimento.



Número de capas ocultas

Realizamos experimentos con 5, 10 y 15 capas ocultas. Entrenamos durante 50 épocas con 256 neuronas por capa y función de activación ReLU. Registramos el valor de la función de error (MSE sobre las 5 predicciones) y del MAE en la mejor época de cada caso.

Número de Capas (N)	Loss (validation)	MAE (validation)
5	2.9196	0.2633
10	5.4508	0.0921
15	8.0909	0.0919

Se observa un peor desempeño al aumentar el tamaño de la red. Según la literatura, esto puede deberse al efecto de gradientes que se desvanece al atravesar una red profunda. Su desempeño se puede mejorar implementando conexiones residuales como el caso de ResNet.

Número de neuronas por capa

Realizamos experimentos con 128, 256 y 512 neuronas por capa. Entrenamos durante 50 épocas con 8 capas ocultas y función de activación ReLU. Registramos el valor de la función de error (MSE sobre las 5 predicciones) y del MAE en la mejor época de cada caso.

Número de Neuronas (K)	Loss (validation)	MAE (validation)
128	3.7015	0.2270
256	4.3431	0.1141
512	3.7988	0.0917

Los resultados no son concluyentes. Sin embargo, se observa un degrado en el rendimiento para 256 neuronas por capa comparado con el rendimiento de los otros 2 casos evaluados.

Función de activación

Realizamos experimentos con funciones de activación ReLU, Tangente Hiperbólica y Sigmoide. Entrenamos durante 50 épocas con 12 capas ocultas y 64 neuronas por capa. Registramos el valor de la función de error (MSE sobre las 5 predicciones) y del MAE en la mejor época.

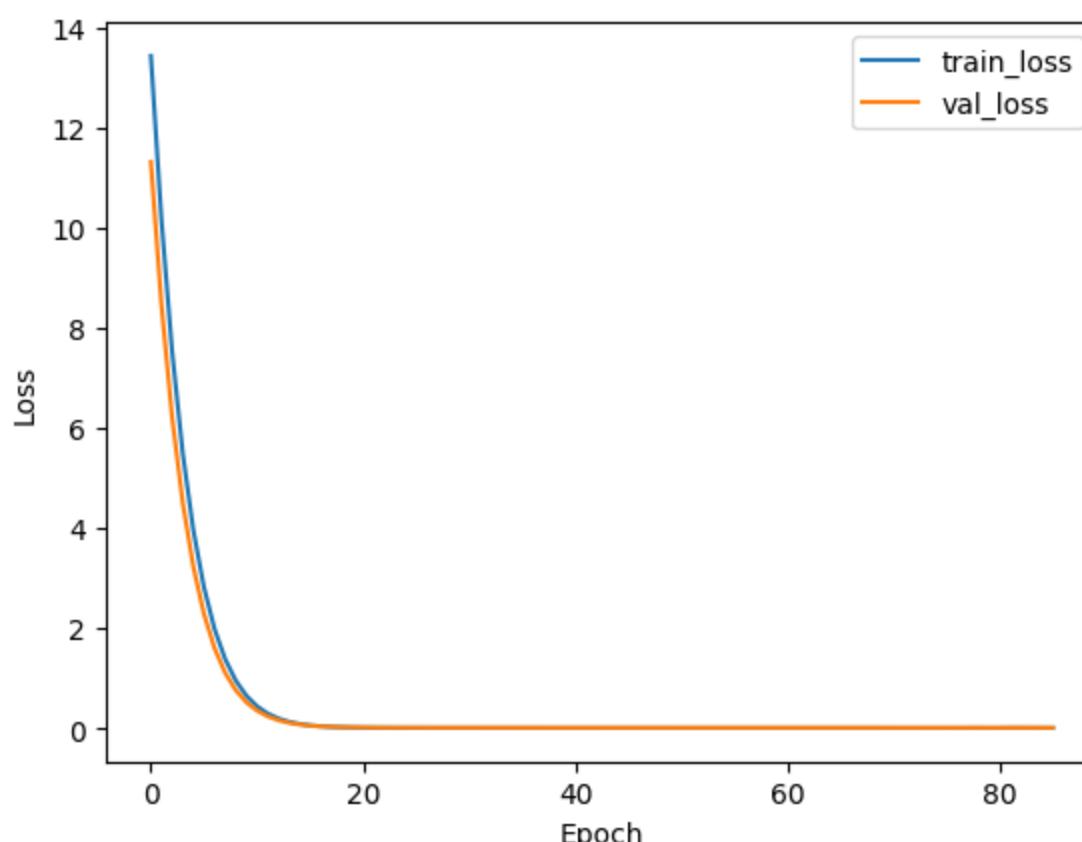
Función de Activación (F)	Loss (validation)	MAE (validation)
ReLU	3.8428	0.1271
tanh	3.6831	0.0924
sigmoid	3.6950	0.1143

Se observan resultados estadísticamente similares para las tres funciones de activación. Las variaciones observadas pueden atribuirse a factores externos más allá de la utilidad implícita de cada una. Por conveniencia y rapidez en su cálculo, la literatura recomienda utilizar ReLU.

Modelo seleccionado

A partir de los hallazgos de los experimentos, se decidió escoger 5 capas ocultas con 256 neuronas cada una y función de activación ReLU. La función de activación se escoge ya que obtiene resultados similares a las demás funciones de activación, con un menor coste computacional; por su parte, la combinación de capas ocultas y número de neuronas demostró ser útil para evitar un entrenamiento demasiado lento del modelo.

Los experimentos además demostraron que otro hiperparámetro es clave para el rendimiento del modelo: la cantidad de épocas. En los experimentos realizados, el modelo terminaba su entrenamiento sin haber llegado nunca a un punto en el cual su rendimiento parara de mejorar, lo que sugiere que estaba siendo sub-entrenado para las capacidades posibles. Por este motivo, en el modelo final se decidió entrenar durante 250 épocas con un callback de Early Stopping después de 10 épocas sin mejoras. El modelo se entrenó hasta la época 76 y alcanzó un desempeño con un valor en validación de la función de error del 0.0131 y un MAP del 0.0897, los cuales se encuentran muy por debajo del mejor modelo encontrado en los experimentos anteriores. De igual forma, al observar su curva de entrenamiento, se observa que alcanzó el estado de plateau característico de los retornos decrecientes del proceso de entrenamiento.

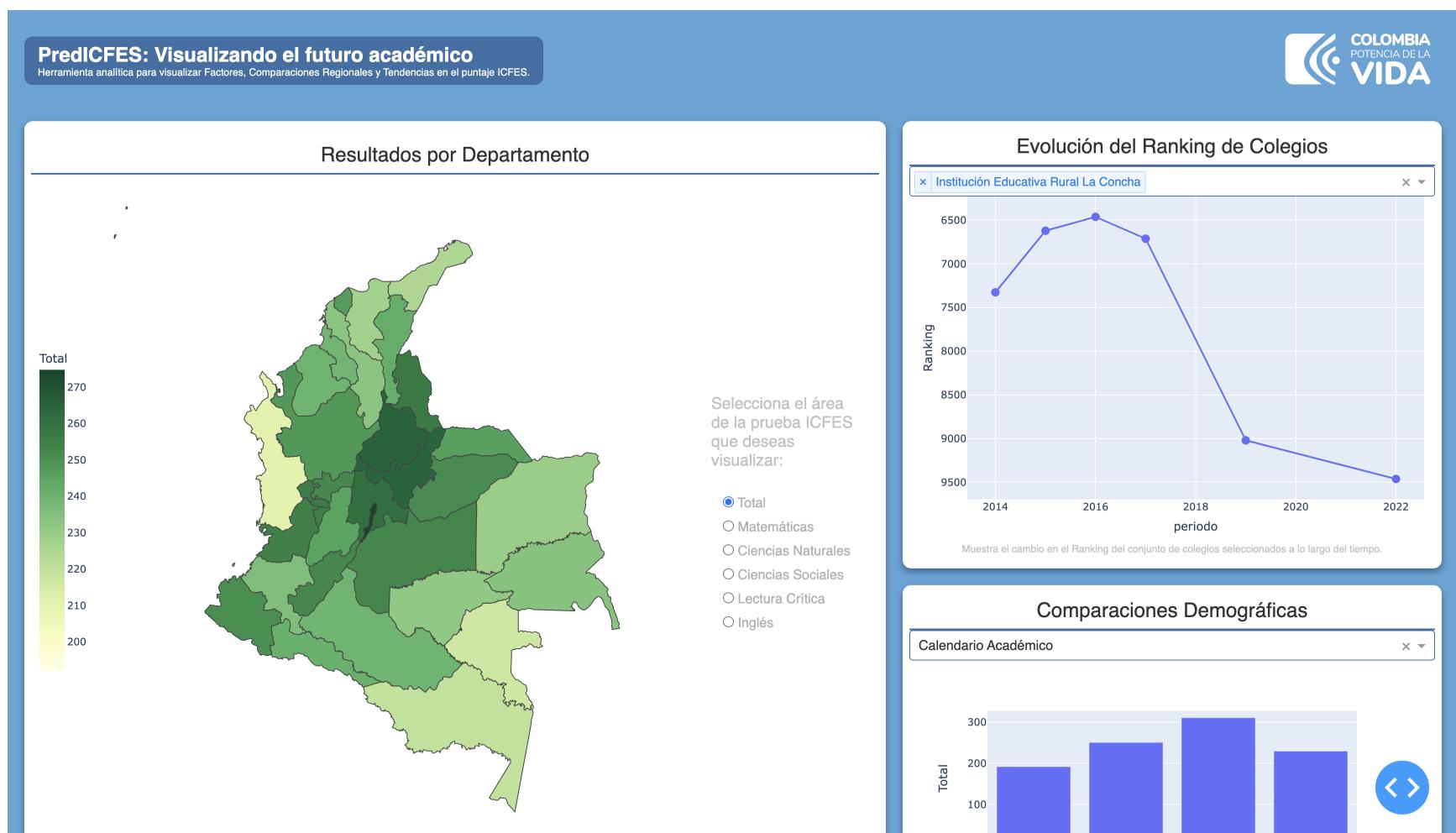


Dashboard

PredICFES cuenta con 4 funcionalidades principales:

- Identificar a través de un mapa los departamentos con mejores y peores desempeños.
- Hacer un seguimiento al *ranking* nacional obtenido por los colegios colombianos.
- Visualizar el desempeño promedio de diversos grupos demográficos en las pruebas de estado.

- Predecir el puntaje obtenido por un estudiante en las distintas competencias y de manera global en el examen ICFES Saber 11.



Disponible en: <http://52.54.224.49>