Predicción de Resultados en Pruebas Saber Pro a partir de Saber 11 utilizando Modelos de Machine Learning

Jaime Arturo Ramírez Cortes
Juan Pablo Vera Fuentes
Sebastián Barrera Bernal
Mateo Grisales Hurtado

Introducción

Este proyecto tiene como objetivo principal desarrollar un modelo predictivo que permita anticipar los resultados de las pruebas Saber Pro basándose en la información disponible de las pruebas Saber 11. A partir de un análisis profundo de datos históricos, el proyecto busca identificar patrones y factores relevantes que contribuyan al desempeño de los estudiantes en educación superior. Utilizando técnicas de machine learning, este modelo pretende ser una herramienta útil para instituciones educativas, ayudándolas a reconocer áreas de mejora y posibles intervenciones en etapas tempranas de formación académica.

Definición del Problema y Pregunta de Negocio

Problema: Existe una necesidad de entender cómo las condiciones académicas y socioeconómicas de los estudiantes al finalizar su educación secundaria (Saber 11) pueden influir en su rendimiento en la educación superior (Saber Pro). Este proyecto busca responder a esta necesidad mediante la identificación de factores predictivos en los datos de Saber 11 que se relacionen directamente con los resultados de Saber Pro, permitiendo a las instituciones educativas desarrollar estrategias de mejora basadas en datos.

Pregunta de negocio: ¿Cómo podemos utilizar los datos de Saber 11 para predecir los resultados en Saber Pro y, en consecuencia, implementar intervenciones educativas que mejoren el rendimiento en la educación superior?

Descripción y Exploración de los Datos

Para este proyecto, se han utilizado datos provenientes de las pruebas Saber 11 (2010-2022) y Saber Pro (2018-2022) obtenidos del portal de datos abiertos del gobierno colombiano. Ambos conjuntos de datos contienen variables clave, como la ubicación geográfica, el estrato socioeconómico, las condiciones de vivienda y los resultados de

los exámenes, con un total de 51 y 57 columnas respectivamente. Describiremos brevemente las variables principales:

Variables Categóricas

En primer lugar, se identificaron varias variables que contienen un solo valor o presentan una alta concentración en una categoría. Es el caso de ESTU_ESTUDIANTE, ESTU_NACIONALIDAD y ESTU_PAIS_RESIDE, las cuales serán eliminadas del análisis. Asimismo, la variable COLE_BILINGUE muestra una marcada predominancia del valor "No", lo que refleja una tendencia esperada en el contexto colombiano.

Otro hallazgo importante es el desbalance en las frecuencias de la variable INST_ORIGEN en Saber Pro, que presenta una mayor proporción de categorías como "NO OFICIAL - CORPORACIÓN" y "NO OFICIAL - FUNDACIÓN". Para simplificar el análisis, se recomienda agrupar estas últimas en una categoría única denominada "NO OFICIAL".

Además, se detectaron inconsistencias en los nombres de algunos municipios, como BOGOTÁ, que aparecen tanto en mayúsculas como en minúsculas. Es fundamental estandarizar estas nomenclaturas para evitar duplicaciones y errores en el análisis. Este mismo enfoque debe aplicarse a otras variables relacionadas con la ubicación en ambos conjuntos de datos.

Por otra parte, se observó que ciertas variables categóricas, como COLE_COD_DANE_ESTABLECIMIENTO y ESTU_PRGM_ACADEMICO, tienen una cantidad extensa de valores únicos, lo que podría complicar el análisis. Se sugiere considerar la agrupación de ciertos códigos o especializaciones que resulten muy específicos y cuenten con pocas observaciones.

En cuanto a las variables socioeconómicas, FAMI_ESTRATOVIVIENDA y FAMI_EDUCACIONPADRE/MADRE presentan una representación adecuada de las categorías, lo cual es favorable para el análisis posterior. Asimismo, variables relacionadas con características del hogar, como FAMI_TIENEAUTOMOVIL, FAMI_TIENECOMPUTADOR, FAMI_TIENEINTERNET y FAMI_TIENELAVADORA, están balanceadas, lo que mejora la calidad del análisis y del modelado.

Distribución de Variables Numéricas

El análisis de las variables numéricas muestra una amplia variabilidad en la mayoría de los casos, lo que permite capturar distintas características de la población estudiada. La variable EDAD_EXAMEN presenta una distribución razonable para el contexto de las pruebas, aunque se recomienda analizar la mediana y la moda para identificar posibles estudiantes atípicos que presenten las pruebas a edades no convencionales.

En las variables socioeconómicas de Saber Pro, se observa una distribución sesgada hacia categorías más bajas, lo que podría corroborar un contexto socioeconómico predominante en la muestra. Esta información es crucial para evaluar el impacto de los factores socioeconómicos en los resultados de las pruebas.

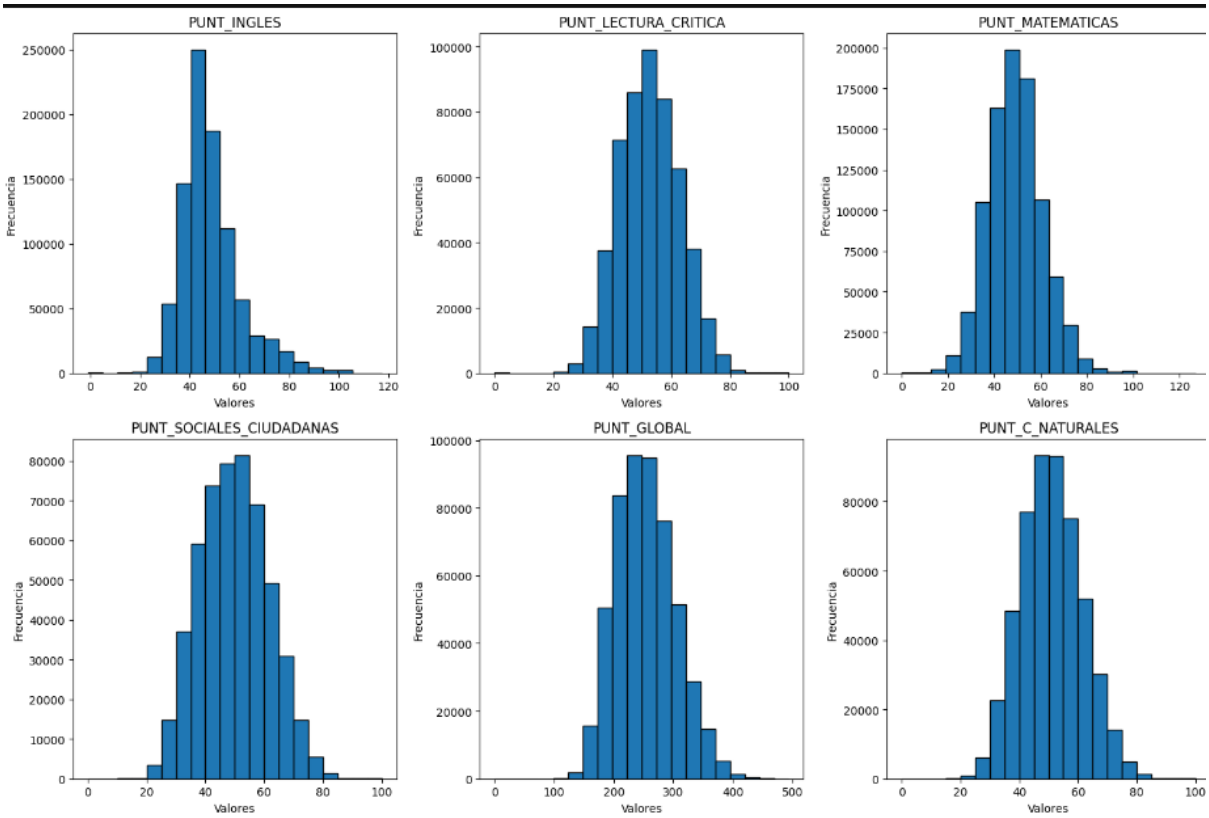
Respecto a los puntajes, se identificó que la distribución de los mismos varía entre áreas, mostrando tanto una curva de distribución normal como sesgos en ciertos casos. Este comportamiento sugiere diferencias en la dificultad de las pruebas o en el rendimiento de los estudiantes. Por lo tanto, se recomienda la normalización o el escalado de estos valores en futuras fases del proyecto.

En términos de representación geográfica, las variables se han tratado como categorías, lo cual es apropiado. Sin embargo, se debe evaluar cómo abordar las regiones con alta concentración de datos y considerar la agrupación de las ubicaciones para facilitar las comparaciones.

Por último, es importante destacar que los puntajes de las pruebas Saber no siguen una escala tradicional de calificaciones, sino que están diseñados para permitir comparaciones entre estudiantes de diferentes contextos. Esto ofrece una perspectiva diferente sobre los resultados.

A continuación, mostraremos la distribución de las variables numéricas principales que encontramos en las bases de datos:

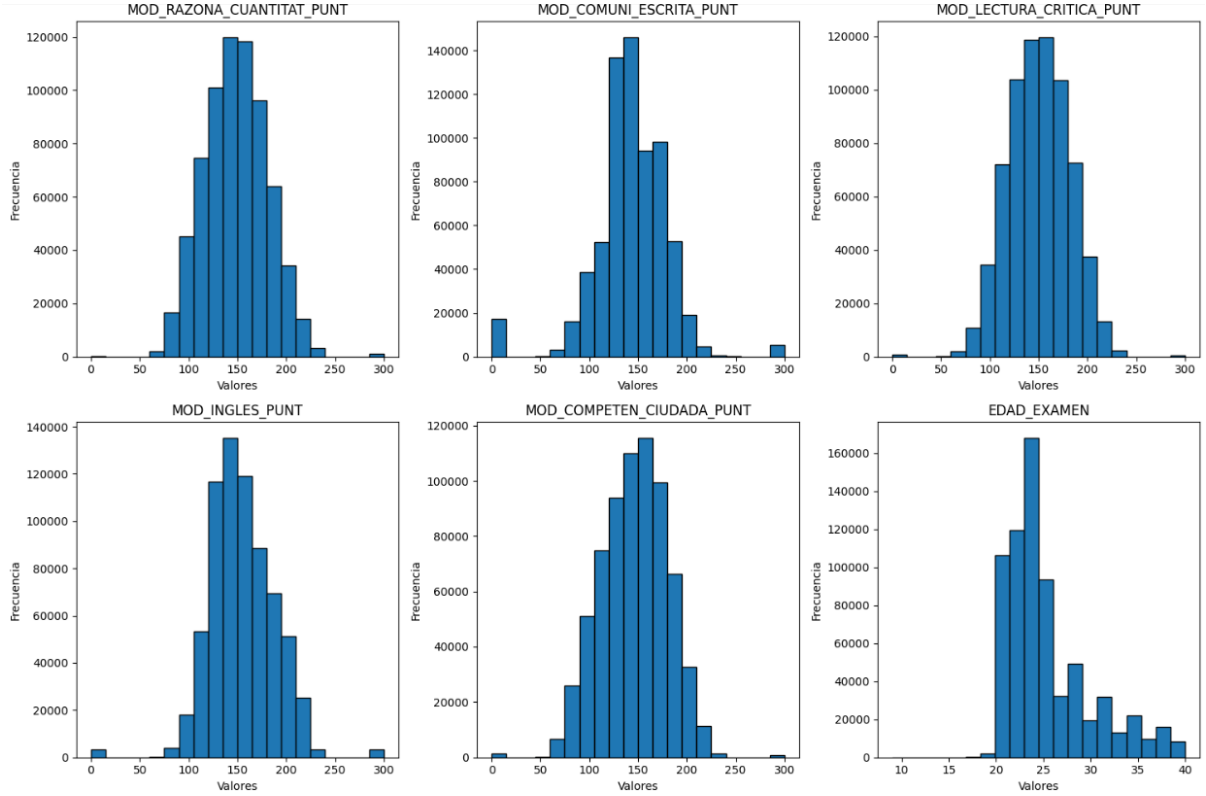
Saber 11:



En la prueba Saber 11, se observa que las variables que representan los puntajes en cada una de las áreas de evaluación presentan una distribución normal. Las medias de las pruebas son las siguientes: ciencias sociales y ciudadanas (49.2), ciencias naturales (50.4), lectura crítica (52.0), matemáticas (49.2) e inglés (48.4). Estos resultados indican que la media de los estudiantes que presentaron esta prueba alcanzó aproximadamente el 50% de respuestas correctas. Esta tendencia se refleja también en el puntaje global, que tiene una media de 253.8 sobre un total de 500 puntos posibles.

La prueba es generalmente completada por estudiantes, lo que resulta en una distribución de edad que se concentra entre los 16 y 20 años, con una media de 17 años. Aunque se registran algunos valores de edad superiores a 20 años, estos son considerados atípicos en el contexto de la distribución.

Saber Pro:



Aquí tienes la redacción revisada y mejorada:

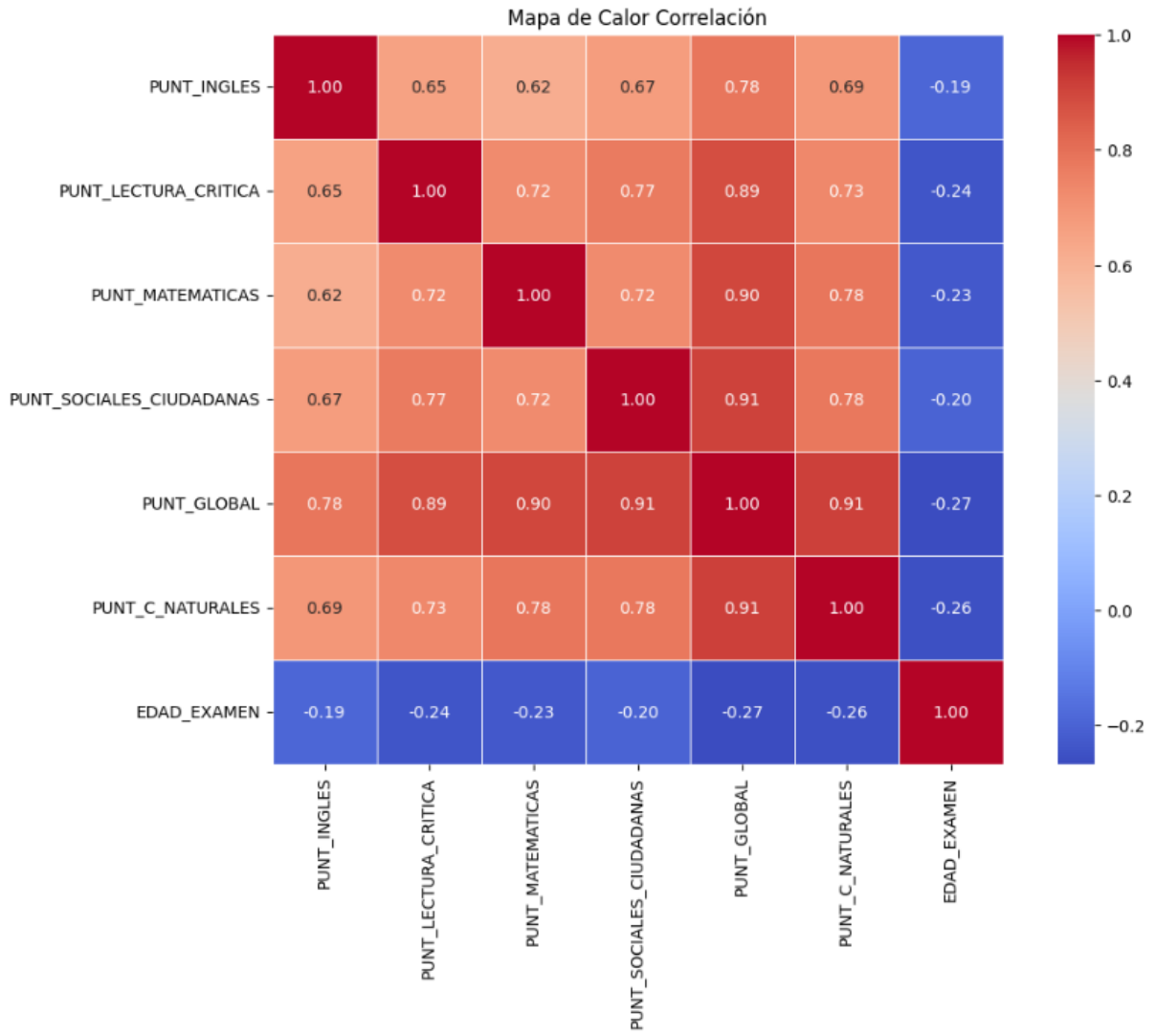
Para la prueba Saber Pro, es interesante observar que la distribución de los resultados difiere notablemente de la presentada en el Saber 11. Los análisis revelan la presencia de más valores atípicos en los resultados de esta prueba, y aunque se conserva una distribución que parece normal, se observa que en algunas áreas la cola se inclina hacia la izquierda o hacia la derecha.

Aunque las pruebas evalúan habilidades similares, acorde al nivel educativo alcanzado por los estudiantes, no se cuenta con una variable de puntaje global en el Saber Pro. Cada área del conocimiento se califica en una escala de 0 a 300. Es relevante destacar que las medias de desempeño en cada área del Saber Pro son comparables a las del Saber 11: razonamiento cuantitativo (148), lectura crítica (149.9), comunicación escrita (142.1), inglés (154.3) y competencias ciudadanas (145.6).

Además, la media de edad de los estudiantes que presentan esta prueba es de 25 años, aunque se observa una predominancia de la cola derecha en la distribución, con valores significativos que se extienden hasta los 40 años.

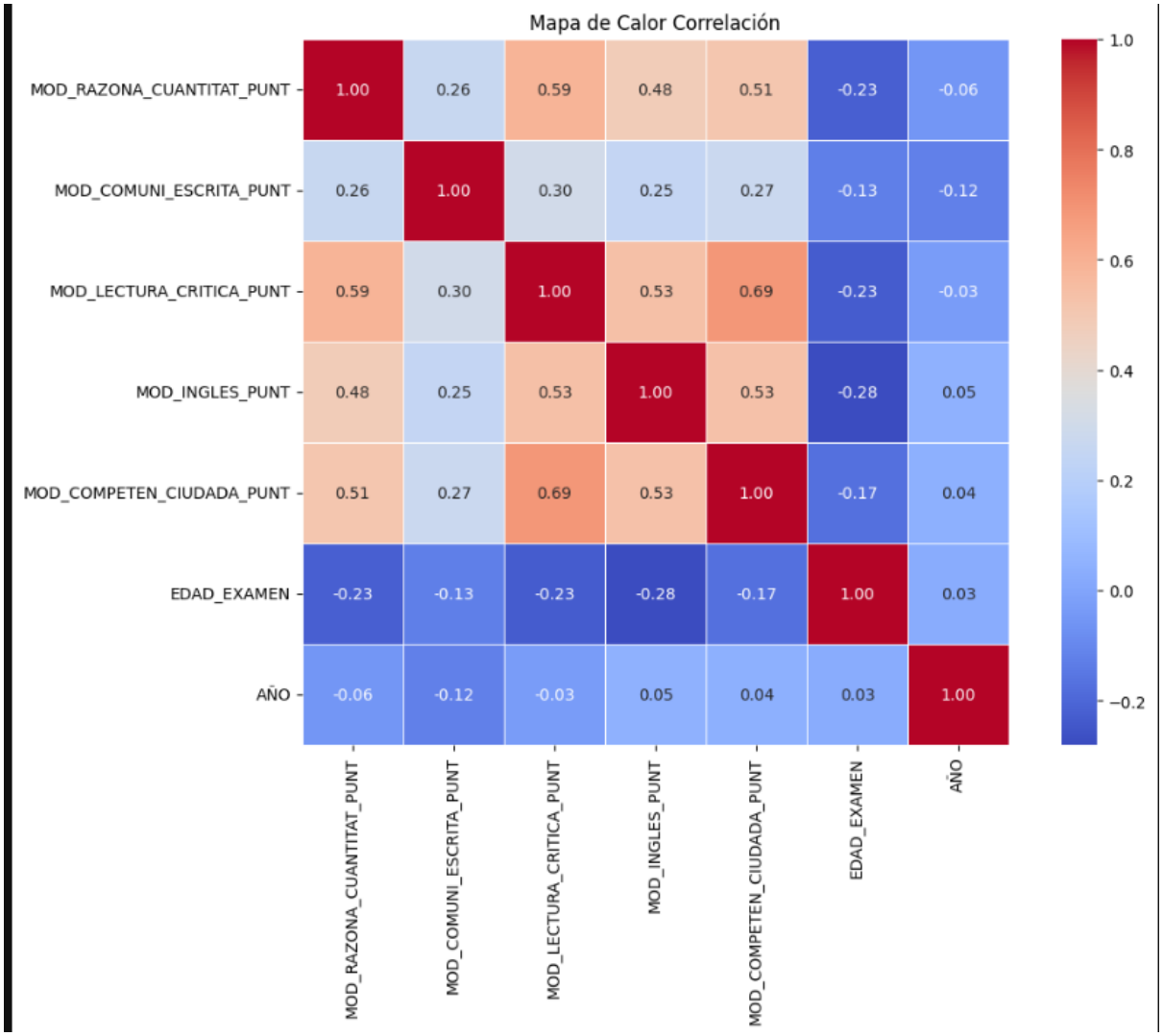
En cuanto a la Correlación entre las variables encontramos lo siguiente:

Saber 11:



En las variables de interés, encontramos correlaciones significativas entre los puntajes, lo que sugiere que un estudiante tiende a obtener resultados similares en todas las pruebas que presenta. Este fenómeno puede explicarse por las condiciones socioeconómicas de los estudiantes, donde los resultados dependen de la calidad del conocimiento recibido previamente. Así, un estudiante se desempeñará de manera similar en cada uno de los campos evaluados. Esta observación respalda las medias estudiadas anteriormente, ya que, en promedio, un estudiante obtendrá resultados bastante homogéneos en cada prueba que presente, manteniendo una media similar entre todas ellas.

Saber Pro:



A diferencia de la prueba Saber 11, la correlación de los puntajes en la prueba Saber Pro disminuye. Esta prueba es presentada por estudiantes universitarios que buscan finalizar sus estudios, lo que implica que han desarrollado habilidades diferentes durante su formación académica. Por lo tanto, tiene sentido que su desempeño en distintas áreas del conocimiento empiece a variar. Un estudiante que adquiere conocimientos generales en diversas disciplinas no se comporta igual que un profesional especializado en un área específica.

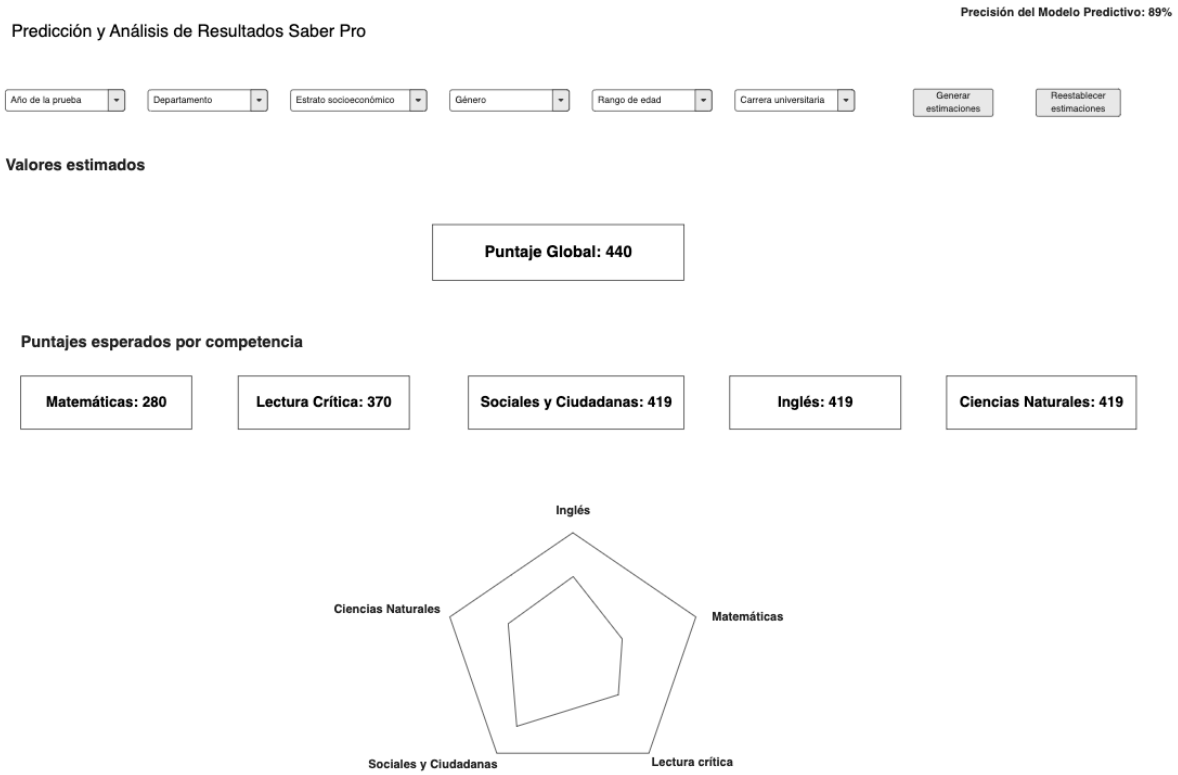
Para el desarrollo del proyecto, es relevante identificar qué factores de la prueba Saber 11 pueden influir en el rendimiento de la prueba Saber Pro. Aunque la tendencia hacia la media se mantiene en el 50% de los puntos posibles, la correlación de los resultados en cada área específica no es la misma.

Maqueta Inicial del Prototipo

Conforme al cumplimiento de los objetivos planteados para el aplicativo final que apoyará la toma de decisiones en el contexto de las pruebas Saber Pro. Se propone la

creación de un dashboard que tendrá como entradas variables demográficas previamente seleccionadas (estas deberán tener un impacto significativo para todas las carreras involucradas en el estudio). Seguidamente el aplicativo se encargará de presentar visualizaciones que informarán sobre la estimación del modelo para el rendimiento del sujeto en las diferentes áreas de conocimiento. Adicionalmente se presentarán visualizaciones para comparar al estudiante con los puntajes globales de los estudiantes de esa disciplina en las diferentes áreas de conocimiento.

Finalmente, basado en unos intervalos definidos, se generarán recomendaciones para el estudiante considerando el rendimiento estimado para cada área.



Promedio globales para resultados especificos

Recomendaciones

Debido a que el puntaje de matemáticas se encuentra bajo (250 - 300), se sugiere reforzar este campo de conocimiento.

Repositorio y Contenido

El repositorio del proyecto se encuentra en GitHub en el siguiente enlace: Predicción de Resultados Saber Pro. En este repositorio se almacena toda la información relevante para el proyecto, organizada en diferentes carpetas para facilitar el acceso y la colaboración entre los miembros del equipo. A continuación, se presenta la estructura del repositorio y la descripción de cada una de sus carpetas:

1. **Datos y Directorios de Datos**:
    a. Esta carpeta contiene enlaces y archivos relacionados con los datos originales de las pruebas Saber 11 y Saber Pro. Incluye:
        i. `Link_Datos_Saber_11_Originales.txt`: Un archivo con el enlace de descarga para los datos originales de Saber 11.
        ii. `Metadatos Saber 11 y Saber Pro.xlsx`: Un archivo de Excel con la descripción y metadatos de los conjuntos de datos.
        iii. `Resultados_Saber_Pro_Original.zip`: Archivo comprimido con los datos originales de Saber Pro.
2. **Documentos**:
    a. En esta carpeta se almacenan los documentos de planeación y manejo de datos del proyecto. Actualmente contiene:
        i. `Entrega_1_plan_manejo_datos.pdf`: Documento que detalla el plan de manejo de datos, incluyendo aspectos éticos, legales y metodológicos relacionados con el uso y almacenamiento de los datos.
3. **Versiones de Código**:
    a. Aquí se encuentra el código utilizado para la exploración y limpieza de los datos, documentado en el notebook:

i. `Análisis_Saber11_SaberPro.ipynb`: Un notebook que contiene el proceso de análisis y limpieza de datos realizado sobre las pruebas Saber 11 y Saber Pro.

Cada una de estas carpetas ha sido configurada para el control de versiones mediante Git, permitiendo un manejo adecuado tanto del código como de los datos y facilitando la trazabilidad de cambios y la colaboración en equipo.

Documentación y Versionado

Para asegurar la trazabilidad y control de versiones, se está utilizando Git para el código y DVC para el manejo de datos. El repositorio de GitHub contiene el código y los scripts de limpieza y análisis, mientras que los archivos de datos originales y limpios están versionados con DVC, lo que permite al equipo seguir los cambios realizados a lo largo del proyecto. Además, se han organizado las carpetas del repositorio para facilitar el acceso a los datos y metadatos por parte de todos los miembros del equipo.