**Reporte del proyecto: “Desarrollo de una herramienta de predicción del Saber 11 basada en Modelos de Inferencia Bayesiana”**

Sebastián David Acero - 202014058

Sara María Troncoso - 202013766

Pedro Cedeño – 201820892

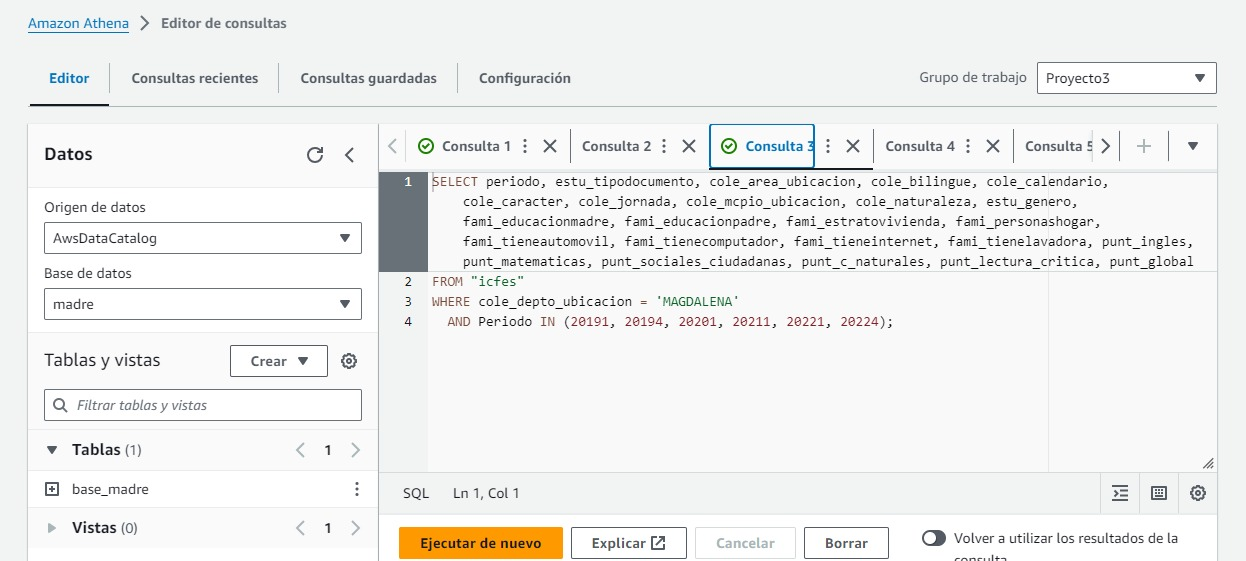
**Introducción**

El propósito de este documento es presentar el diseño de un proyecto cuyo objetivo principal es desarrollar una herramienta capaz de predecir la probabilidad de obtener un resultado sobresaliente en el examen del Saber 11 y mostrar información de interés general por medio de visualizaciones. Para lograr este objetivo, se llevará a cabo la implementación de un modelo de inferencia bayesiana que se fundamentará en una base de datos “Resultados únicos Saber 11” suministrada por el Instituto colombiano para la evaluación de la educación. Esta base de datos, actualizada por última vez el 21 de noviembre de 2023 contiene alrededor de 7 millones de vistas e información acerca de los examinados, las cualidades de su colegio, el departamento en el que está ubicado, el género y gran cantidad de datos necesarios para lograr esta inferencia. Este producto de analítica de datos está diseñado para ciudadanos interesados en conocer el estado de la educación específicamente en el departamento del Magdalena.

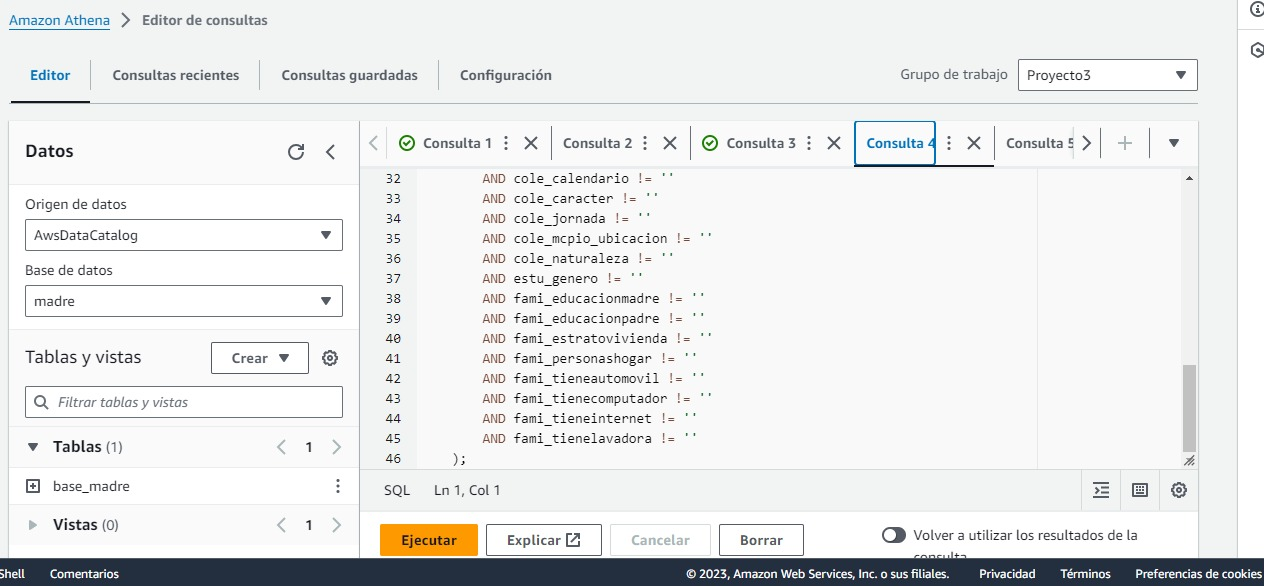
**Exploración de Datos**

A continuación, se llevará a cabo un análisis más exhaustivo de la base de datos con el propósito de llevar a cabo una limpieza de los datos, seleccionando únicamente aquellos que resulten más pertinentes para la evaluación. Específicamente, se decidió trabajar con los datos del departamento del Magdalena ya que desde hace un tiempo su gobernación ha concentrado sus esfuerzos en la “Revolución de la educación” según el Plan de Desarrollo Magdalena Renace (Gobernación del Magdalena, 2023). Esto causó gran interés en nosotros debido a que en un futuro queremos ver como este mejoramiento en la infraestructura educativa, la dotación escolar y la concentración del Programa de Alimentación Escolar tiene un resultado positivo en el resultado de la educación y como esto se verá reflejado en los resultados del Saber 11. Nuestro objetivo es en unos años actualizar la base de datos y ver como ha sido esta diferencia. También, quisimos observar como afectó la pandemia en los resultados de esta prueba.

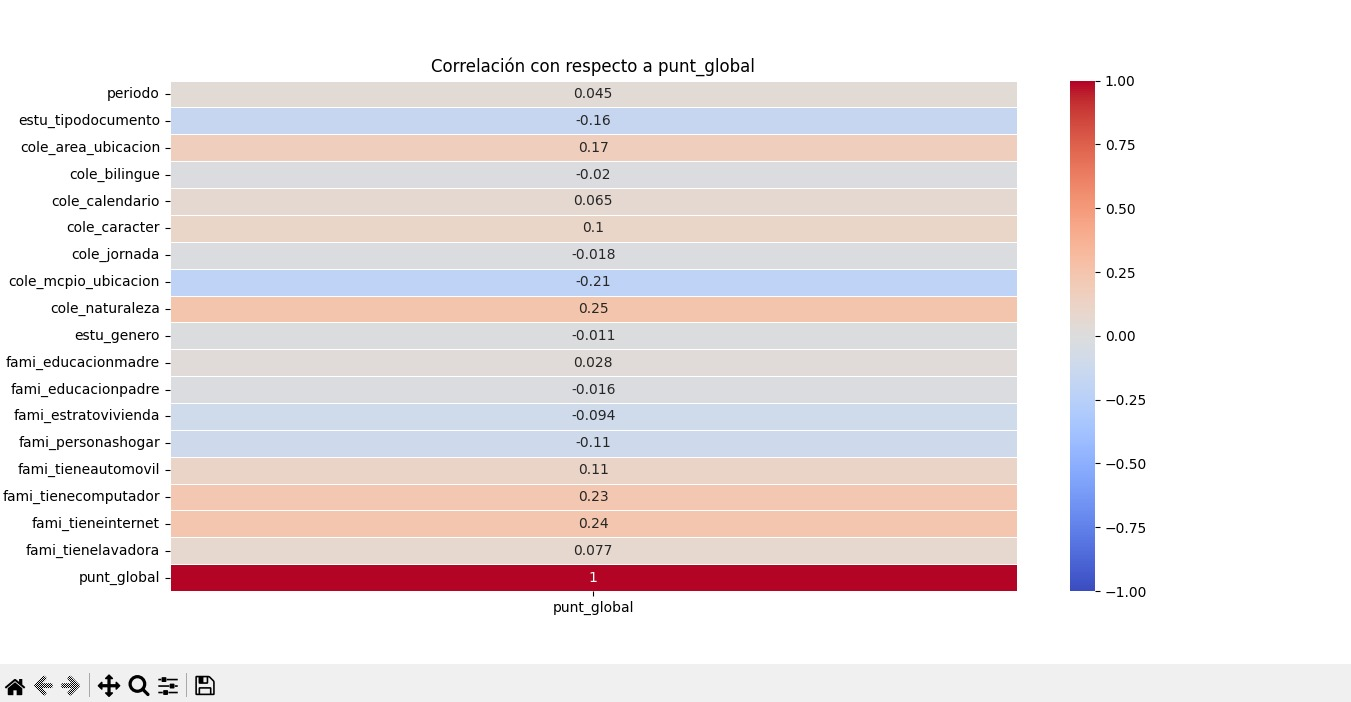
Dentro de nuestra base de datos, tomamos decisiones basadas en análisis estadísticos que nos ayudaron a identificar cuales eran las variables más relevantes para alcanzar los objetivos del proyecto. En primera instancia, dado que el objetivo principal de este proyecto era evaluar la probabilidad que tienen los estudiantes de obtener un resultado sobresaliente en el examen la variable Puntaje Global era de vital importancia en el marco del proyecto. Para poder filtrar los datos, utilizamos AWS Athenea para extraer los subconjuntos de datos relevantes, en este caso filtramos la ubicación del colegio por departamento la cual estrictamente tenía que ser Magdalena y específicamente quisimos observar los datos de 2019 a 2022 en los cuales ocurrió la pandemia.



Posteriormente, al ver que había una gran cantidad de datos y que en algunos casos había datos vacíos debimos quitar todas las filas que tuvieran celdas vacías con el fin de obtener una base de datos completa.



Para poder mirar la correlación de las variables y así poder escoger con claridad cuales queríamos escoger, utilizamos un mapa de calor el cual es una representación gráfica de los datos donde los valores se representan con colores. Esto fue lo que obtuvimos:

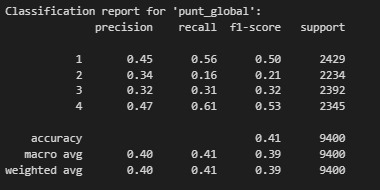


Con la información obtenida tomamos la decisión de solo tener en cuenta aquellas que tengan un valor mayor a 0.1 y menor a -0.1 con el objetivo de identificar las relaciones lineales más fuertes en los datos y así poder filtrar las variables relevantes de las irrelevantes. La única variable que escogimos fuera de este umbral fue el periodo debido a la naturaleza de nuestro proyecto ya que delimitar una fecha era indispensable para ver su comportamiento antes, durante y después de la pandemia y también porque dependiendo del periodo las probabilidades cambiaban.

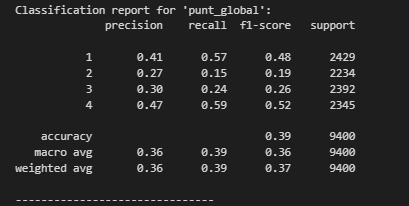
También, para poder analizar los datos fue de vital importancia categorizar cada una de las variables debido a que muchos de ellos contenían información que no era fácil de procesar, por ejemplo, en el caso del puntaje en inglés, matemáticas, competencias sociales y ciudadanas, competencias naturales, lectura crítica y puntaje global fue de gran importancia realizarlo por cuartiles, es decir, valores que dividen una muestra de datos en cuatro partes iguales. Otro caso en el que debimos hacerlo fue en variables que nos proporcionaban si la familia tenia o no computador o internet convirtiéndolos en datos que tomaban valor de 1 si lo tenían y 0 si no.

**Modelación:**

Para poder realizar nuestro modelo de inferencia bayesiana, procedimos a emplear los métodos de aprendizaje de estructura. Inicialmente, implementamos el método "K2", el cual es una técnica que tiene como objetivo la maximización de la probabilidad conjunta de los datos observados, considerando ciertas restricciones de independencia implícitas. El proceso de aplicación de este método incluye las etapas de inicialización, búsqueda, consideración de restricciones, evaluación y, finalmente, selección. Para saber que tan acertado el modelo calculamos las siguientes métricas:

****

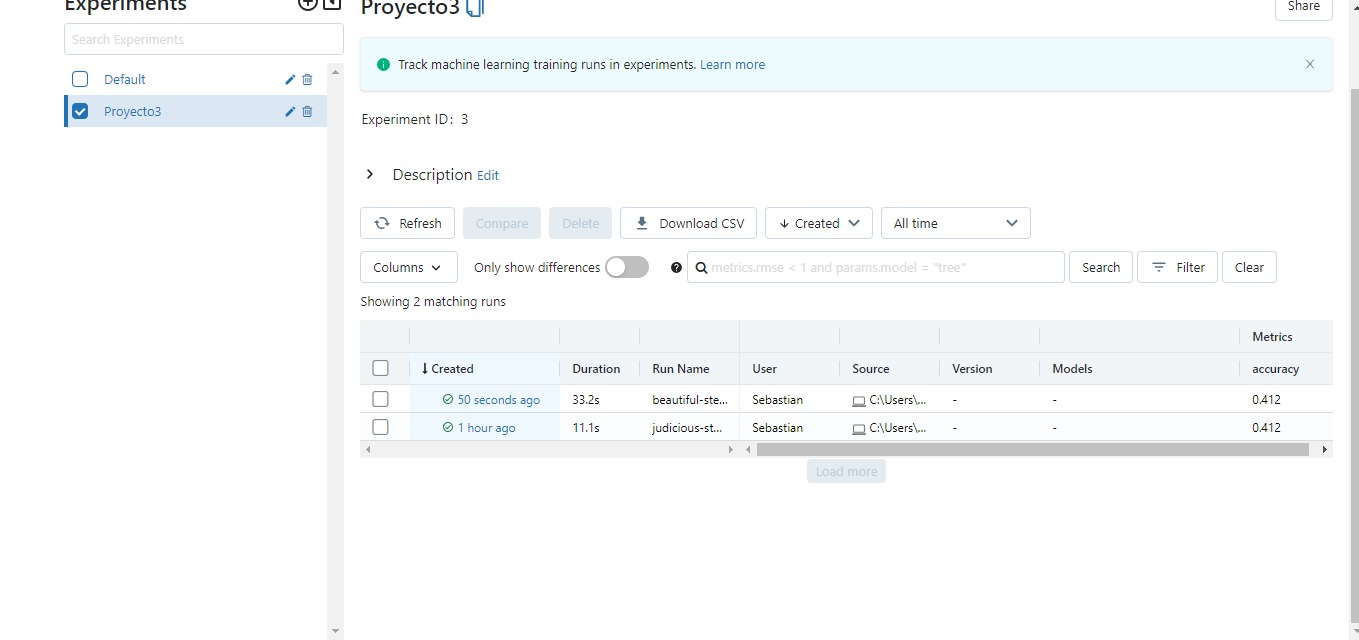
A continuación, procedimos a utilizar el método BIC (Bayesian Information Criterion) Score, el cual tiene como propósito el aprendizaje de la estructura de una red bayesiana a partir de datos observacionales. Este método se emplea para evaluar las diferentes estructuras de red candidatas, buscando identificar aquella que mejor represente los datos observados. Para ello, combina la consideración de la verosimilitud de los datos con la penalización por la complejidad del modelo. El proceso de aplicación del método BIC Score consta de varias etapas, incluyendo una inicialización, una búsqueda activa de estructuras, una evaluación de la idoneidad de estas estructuras y, finalmente, una penalización por complejidad, con el objetivo de seleccionar la estructura de red más apropiada. Nuevamente calculamos las métricas y obtuvimos lo siguiente:



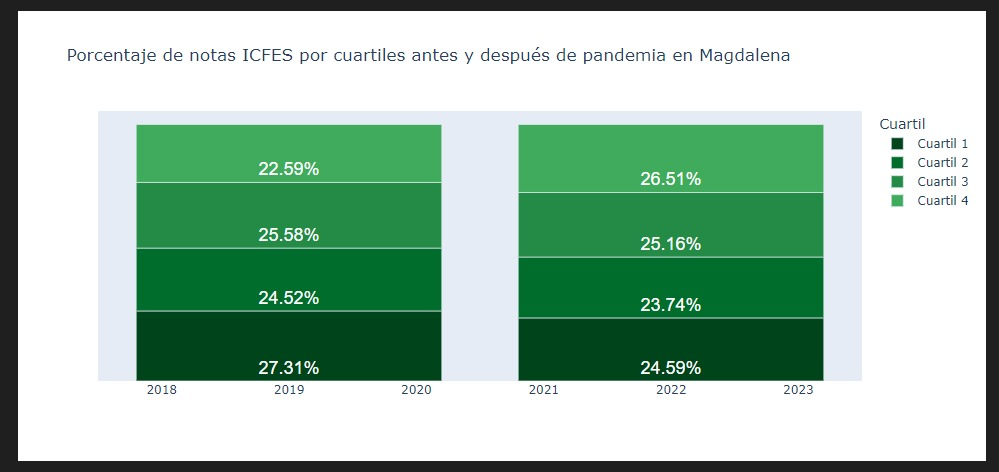
Como podemos ver, evaluamos el rendimiento de los modelos utilizando dos métricas de evaluación esenciales. La primera de estas métricas es el "Recall", también conocido como "Sensibilidad" o "Tasa Positiva Verdadera". Esta métrica se utiliza para medir la capacidad de un modelo para identificar y capturar todos los ejemplos positivos en un conjunto de datos. En otras palabras, busca evaluar la proporción de casos positivos reales que el modelo logra detectar correctamente. La segunda métrica que empleamos fue la "F1 Score". Esta métrica se utiliza en situaciones donde se busca un equilibrio óptimo entre la precisión y el recall. La F1 Score combina la precisión y el recall en una sola métrica, proporcionando una medida más completa del rendimiento de un modelo de clasificación. La tercera métrica fue de support, la cual se utiliza en el contexto de la clasificación y representa el número de instancias o muestras que le pertenecen a una clase específica de un conjunto de datos, conocida también como la frecuencia.

Con estos resultados concluimos que el mejor modelo fue el realizado con el método K2 ya que su precisión fue del 40%, su recall de 41% y su F1 Score de 39% que en comparación con BIC score fue mucho mejor.

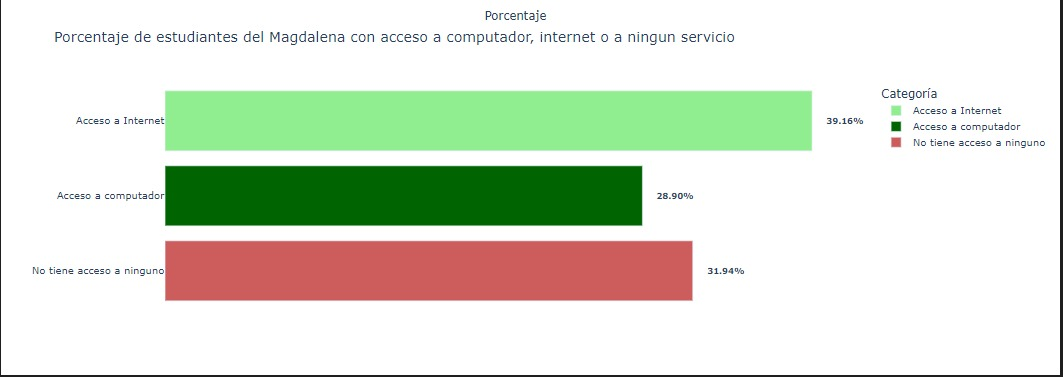
Luego de escoger el modelo, utilizamos MLflow la cual es una plataforma de cógico que esta diseñada para gestionar el ciclo de vida completo del aprendizaje automático con el objetivo de entrenar el modelo y, aunque logramos hacer pruebas, no logramos la forma para probar varios modelos.



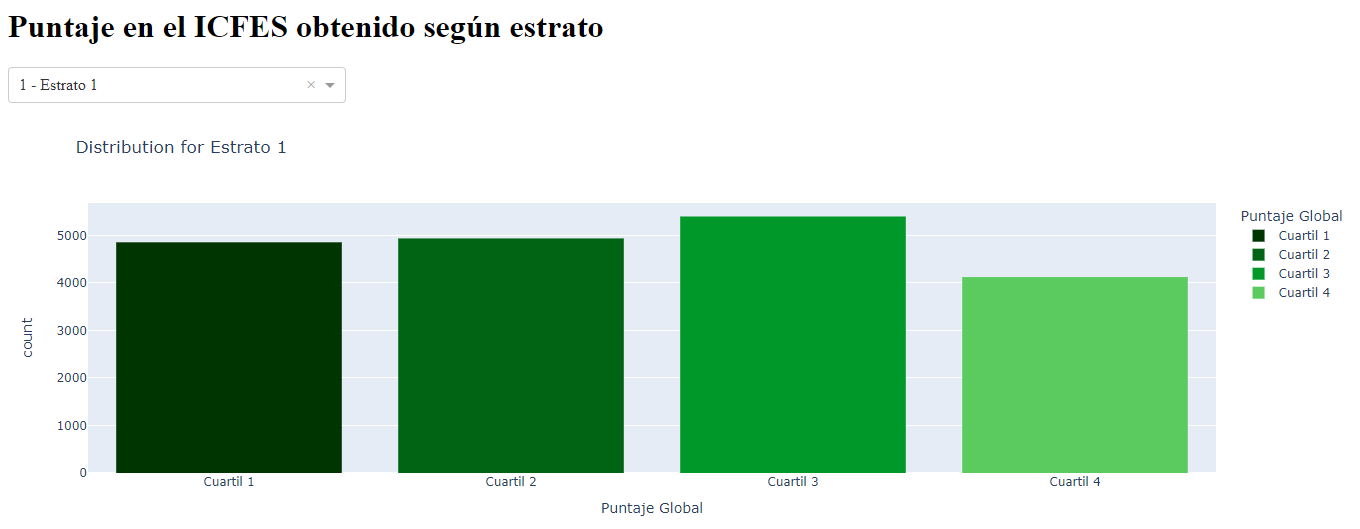
**Visualizaciones:**

****

La primera visualización que realizamos busca mostrar cual es el porcentaje de notas ICFES por cuartiles antes y después de la pandemia en Magdalena. Como podemos ver, en el cuartil 1, 2 y 3 hubo una disminución en el porcentaje de las personas lo que muestra que en 2022 un menor porcentaje de personas obtuvieron notas más bajas en comparación a 2019.



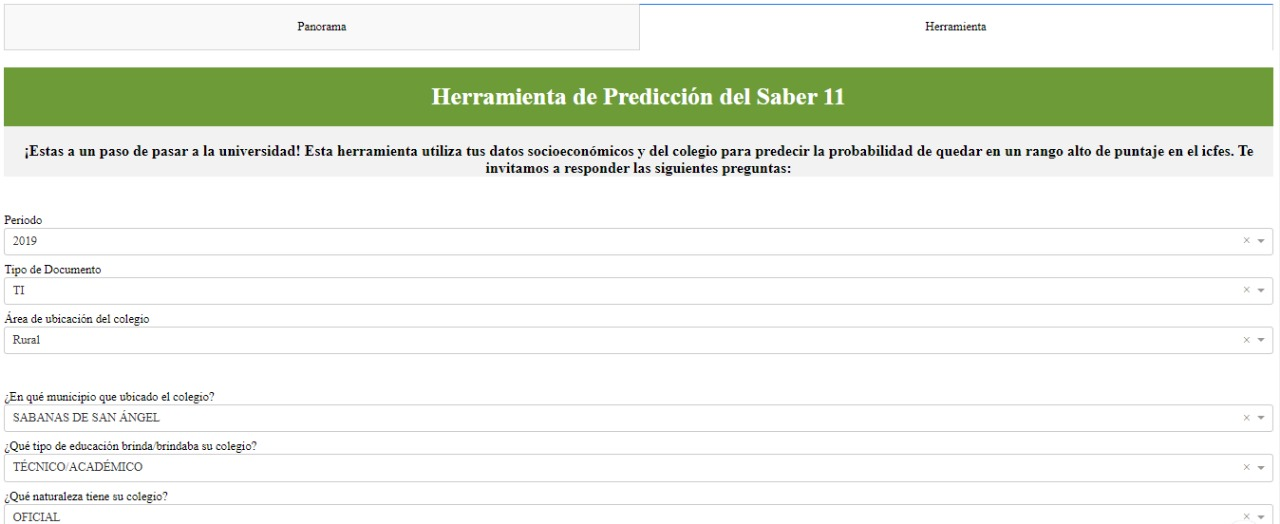
La segunda visualización busca mostrar el porcentaje de personas que tienen acceso a computador, a internet o a ninguno de ellos. Esto lo realizamos con el fin de contextualizar un poco acerca del departamento y podemos ver que menos del 40% de las personas tienen acceso a internet, menos del 30% tiene acceso a computador y que alrededor del 30% no tiene acceso a ninguno. Esto es muy importante debido a que la limitación al acceso de ellos puede implicar una gran dificultad para los estudiantes que en pandemia requirieron educación en línea.

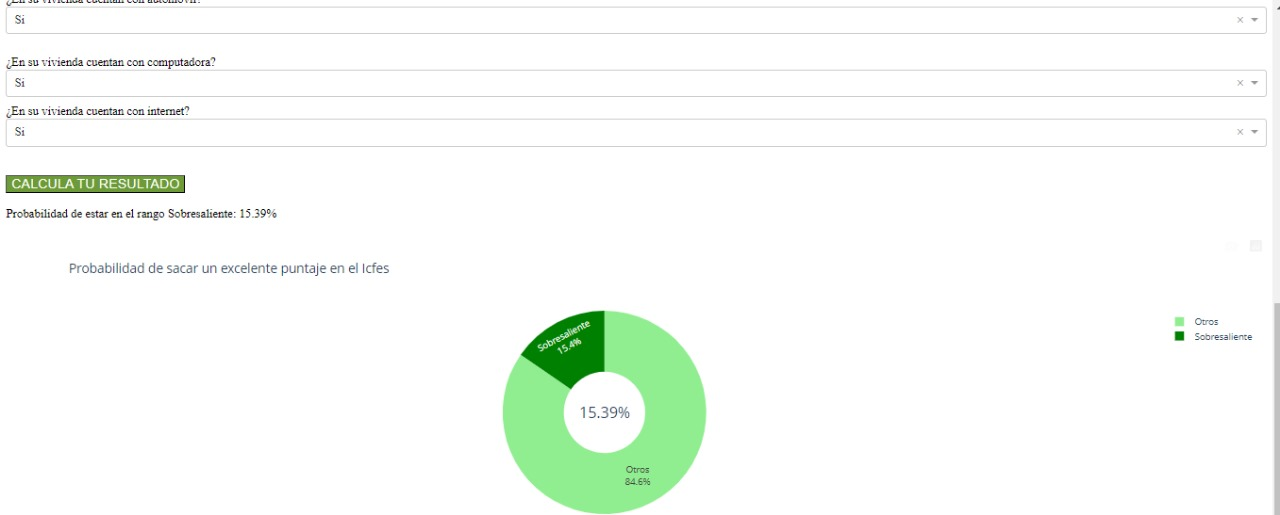


La tercera visualización muestra, por estrato, el puntaje del ICFES obtenido por cuartiles. Como podemos ver, en el caso del Estrato 1 la mayor concentración de personas se encuentra en el cuartil 3 lo cual es bastante positivo.

**Dashboard:**

La construcción de nuestra herramienta consiste en un dashboard, en el que en la primera pestaña se busca mostrar las visualizaciones presentadas anteriormente con el fin de que haya una contextualización del tema. La herramienta se ve de la siguiente manera:



Como se puede ver, la herramienta busca que los estudiantes puedan ingresar sus datos con el fin de predecir la probabilidad de quedar en un rango alto de puntaje en el icfes. La persona debe responder en que periodo está, su tipo de documento, el área de ubicación del colegio, el municipio donde está ubicado, el tipo de educación brindada, la naturaleza de su colegio, si cuentan con acceso a una computadora y por último si tienen acceso a internet. Con esta información calculamos la probabilidad de estar en el rango sobresaliente en el ICFES.

Referencias:

1. Instituto colombiano para la evaluación de la educación (2023). Resultados de las pruebas de Saber 11 del 2010 al 2022. Disponible en: <https://www.datos.gov.co/Educaci-n/Resultados-nicos-Saber-11/kgxf-xxbe/about_data>
2. Gobernación del Magdalena (2023). ¡Así viene cambiando la Educación en el Magdalena! Disponible en: <https://www.gobernaciondelmagdalena.gov.co/asi-viene-cambiando-la-educacion-en-el-magdalena/>