**Analítica Computacional para la toma de Decisiones**

**Proyecto 3 – Predicción Resultados Saber 11**

**Integrantes: Nicolas Ricaurte y Juan Camilo Bello**

**Introducción**

**En nuestro modelo se busca predecir si un estudiante puede o no tener un puntaje superior a 359 dependiendo de sus variables sociales y demográficas.**

**Exploración de los datos relevantes**

En esta parte inicial, para manejar la base de datos del Saber 11 se utilizó AWS Glue y AWS Athena para extraer el subconjunto de datos deseado para evaluar. Denotamos que realizamos 5 Queries usando Athena, para analizar el comportamiento de los datos al igual que escoger la instancia que queríamos trabajar, en el archivo **Exploración inicial** se logra ver estas consultas exactas y como fueron manejadas. En resumen, analizamos:

* Query 1: Promedio de puntaje de acuerdo con la educación de los padres
* Query 2: Promedio de puntaje de acuerdo con el departamento
* Query 3: Promedio del puntaje de acuerdo con el estrato
* Query 4: Cantidad de estudiantes que obtienen un puntaje mayor a 359 de acuerdo con el departamento
* Query 5: Cantidad de estudiantes que obtienen un puntaje mayor a 359 de acuerdo con el estrato

Esto con el fin de resaltar indicadores que considerábamos pertinentes para el modelo que queríamos realizar y más adelante realizar visualizaciones respectivas. Notamos la influencia que tiene la educación de los padres sobre el puntaje global de la prueba, debido a que mediante incrementa el nivel de educación, también los puntajes promediados, esto para los dos padres. Por otro lado, analizamos el comportamiento por departamentos denotando como en las periferias del país se llegan a tener puntajes más bajos en promedio comparado con el centro, donde estos valores incrementan. Por último, buscamos centrar nuestra investigación en las becas que proporciona el estado, para esto buscamos aquellos estudiantes que tuvieran un puntaje de 359 o mayor y los analizamos por departamento y por estrato, encontrando que el departamento con un puntaje promedio más alto es Antioquia, mientras que el estrato dos corresponde al puntaje promedio más alto, indicándonos que es posible que la mayoría de las becas sean otorgadas a personas bajo estas condiciones.

Ahora bien, se decidido analizar el departamento de Cundinamarca para así crear un modelo predictivo alrededor de obtener un puntaje igual o mayor a 359, que corresponde al puntaje necesario para ser becado por el estado, escogiendo variables relevantes para responder esta pregunta de investigación.

Se decide trabajar con el conjunto de datos correspondientes a los periodos superiores al 2019-1 y dentro de Cundinamarca, esto debido a que queremos incluir los fenómenos que se presentaron en la pandemia y cambiaran las dinámicas de estudio de toda la población, mientras que en Cundinamarca es donde se encuentran la mayoría de las entidades de educación superior, de esta forma reduciendo la instancia de la base de datos de forma que sea más fácil manejarla.

Ahora bien, para preparar la base de datos, inicialmente se eliminar se eliminaron las columnas innecesarias, todas aquellas correspondientes a información sobre los colegios puntajes fuera del global y los departamentos, debido a que solo se tratará Cundinamarca. Después, se continuo a realizar las transformaciones correspondientes a cada columna, identificando las variables que realmente correspondían a los datos que necesitábamos, por ejemplo, en la columna de estrato de la vivienda se llevó a cabo una transformación utilizando un mapeo de valores para asignar a cada estrato un número entero del 1 al 6, y asignando el valor 0 para los casos en los que no se tenía información sobre el estrato. Esto mismo se realizó con las demás columnas identificando las variables categóricas y cambiando sus valores.

Se realizaron eliminaciones de columnas innecesarias, lo que resultó en una instancia de datos con las siguientes columnas relevantes: "Genero", "Cuartos en la vivienda", "Educación de la madre", "Educación del padre", "Estrato", "Conviven con cuantas personas", y las variables relacionadas con la disponibilidad de bienes en el hogar, como "Tienen carro", "Tienen computador", "Tienen internet" y "Tienen lavadora". Además, se conservó la columna del puntaje obtenido.

**Modelo de Redes Bayesianas**

Para generar un nuevo modelo se utilizaron métodos de aprendizaje de estructura, se consideraron varios enfoques, incluyendo los métodos basados en restricciones, basados en puntajes y los híbridos. Para este trabajo, se escogió el método de búsqueda de escalada de HillClimb, que es una técnica comúnmente utilizada en la construcción de redes bayesianas.

El método HillClimb comienza con un grafo vacío y se construye iterativamente, añadiendo o eliminando arcos hasta encontrar la mejor estructura que maximiza algún criterio de ajuste, como el puntaje K2. Es importante destacar que durante la implementación del método de HillClimb se puede especificar una lista negra de arcos que se deben evitar durante la construcción del grafo para excluir relaciones conocidas entre variables que se sabe que no existen en el dominio del problema.

Asimismo, se puede establecer un límite en el número máximo de padres para cada nodo para controlar la complejidad del modelo y evitar el sobreajuste. Se buscaron iterar sobre estos parámetros para encontrar el mejor modelo que se ajuste a los datos.

Luego de aplicar el método de HillClimb, se obtuvo un modelo que se ajusta adecuadamente a los datos y se espera que pueda ser utilizado para hacer predicciones y obtener información valiosa. Es importante mencionar que la selección del método de aprendizaje de estructura y la iteración de los parámetros se realizaron con el objetivo de obtener el mejor modelo posible para el conjunto de datos en cuestión, resultando en el modelo de la ilustración 1.

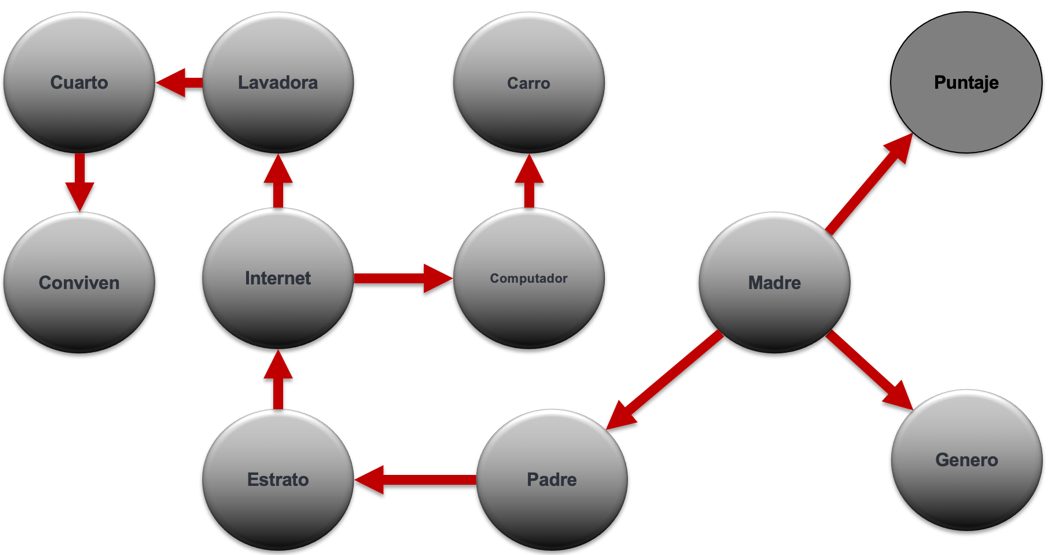


Ilustración 1

La red bayesiana construida para predecir el puntaje de la prueba Saber revela varias relaciones entre las variables estudiadas. Se observa que la presencia de internet en el hogar está relacionada con la posesión de un computador y, a su vez, tener un computador influye en la probabilidad de tener un carro. Asimismo, se encontró que la presencia de internet también está asociada con la presencia de una lavadora en el hogar, y tener una lavadora se relaciona con la cantidad de cuartos en la vivienda. Además, se identificó que la educación de la madre influye en la educación del padre y en el estrato de la vivienda. Por último, la educación de la madre también está relacionada con el género y el puntaje obtenido en la prueba. Estas relaciones observadas en la red bayesiana reflejan patrones y dependencias presentes en los datos analizados, pero es importante recordar que no implican una relación causal directa entre las variables.

**Entrenamiento**

Se utilizó el conjunto de datos " Resultados únicos Saber 11 " del Instituto colombiano para la evaluación de la educación para entrenar y evaluar el rendimiento del modelo bayesiano. Este conjunto de datos contiene información sobre pacientes con problemas cardíacos, y se dividió en dos partes: una para entrenamiento y otra para evaluación. El conjunto de prueba comprendió el 30% de los datos originales, mientras que el conjunto de entrenamiento incluyó el 70% restante.