Sebastián Rodriguez - 201923033

Santiago Torres - 202013737

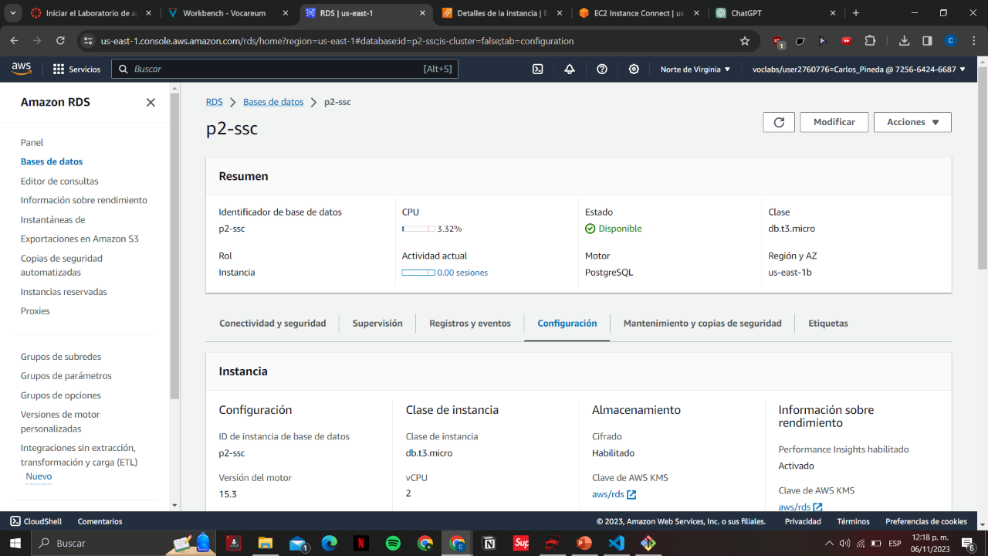
Carlos Pineda – 202013132

## **Proyecto 2 – Predicción del éxito académico v2.0**

**Tarea 1**

En primer lugar, se convirtió el archivo de Excel, donde se contenían los datos de la base de datos que se utilizaron para la red bayesiana del primer proyecto, en un archivo SQL. Esto se hizo mediante un archivo de texto, el cual se comparte dentro de la carpeta de archivos de soporte. A este archivo se le llamo “Estudiantes.sql”.

Después, se creó la Base de datos RDS.



Se conectó la base de datos desde un archivo Python del equipo local.

engine = psycopg2.connect(

    dbname="estudiantes",

    user="postgres",

    password="EstudiantesSSC",

    host="p2-ssc.ccesjqcligmw.us-east-1.rds.amazonaws.com",

    port="5432"

)

cursor = engine.cursor()

Posteriormente, se creó la instancia EC2.

Una captura de pantalla de una computadora

Descripción generada automáticamente

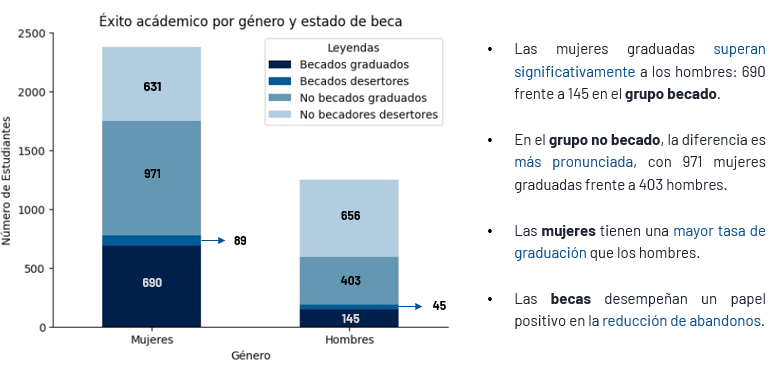
Luego, se copió el archivo SQL en la instancia creada y se llenó una base de datos con los datos de “Estudiantes.sql”.

Captura de pantalla de computadora

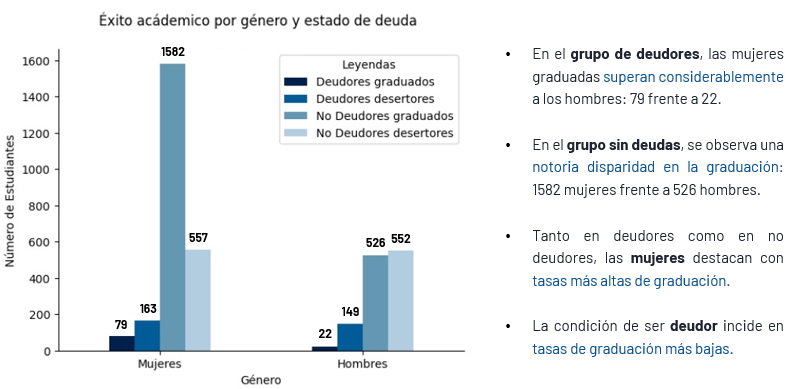
Descripción generada automáticamente

Finalmente, se realizaron las consultas para construir las visualizaciones respectivas.

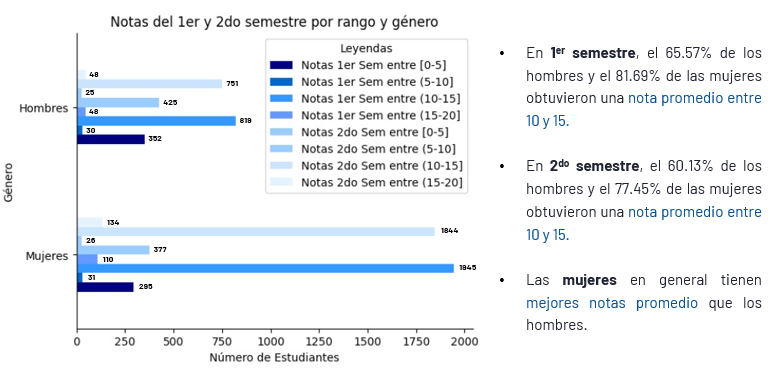
**Primera consulta:** Consulta para obtener el número de estudiantes becados y no becados que se graduaron y desertaron por género.



**Segunda consulta:** Consulta para obtener el número de estudiantes deudores y no deudores que se graduaron y desertaron por género.



**Terca consulta:** Consulta para obtener el número de estudiantes por cada rango de notas promedio y por género.



**Tarea 2**

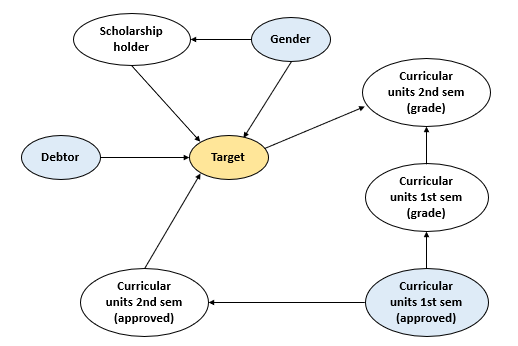
1. Para generar los nuevos modelos que luego compararemos con nuestro modelo del proyecto 1 que fue desarrollado por medio de evidencia bibliográfica de la relación de las variables, utilizaremos dos tres métodos vistos en clase primero modelo por restricciones luego por Hill\_climbing y finalmente por esth\_bic.

**Modelo por restricción:**



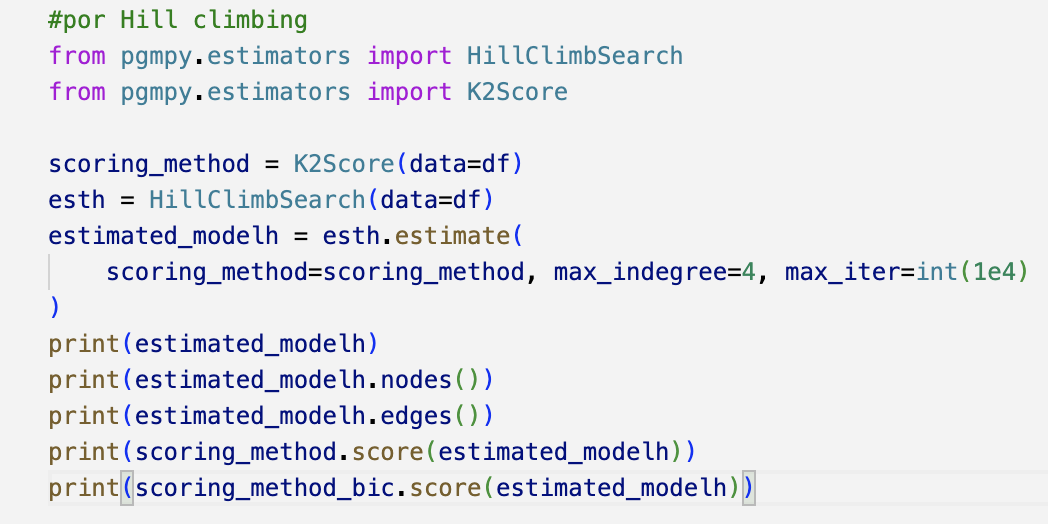
Dándonos un resultado que para este modelo el número máximo de variables condicionantes permitidas es de 4.

Red Bayesiana del modelo:



**Puntaje BIC modelo =** -14313.88035

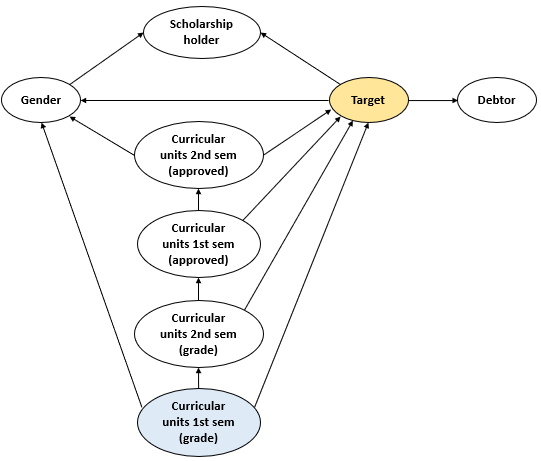
**Modelo por Hill Climbing:**

****

max\_indegree: Es el número máximo de nodos que pueden ser padres de un nodo en la estructura de la red bayesiana en este caso el máximo número de padres que puede tener un nodo es de 4.

max\_iter: Es el número máximo de iteraciones que (HillClimb) realizará, en este caso int(1e4) es una notación abreviada para 10,000 iteraciones.

Red Bayesiana del modelo:



**Puntaje BIC modelo =** -15070.055850

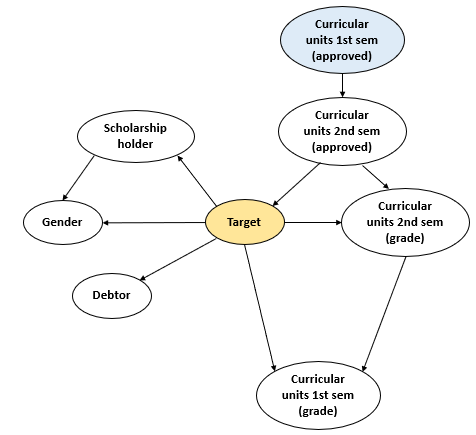
**Scoring\_method\_BIC:**

****

max\_indegree: Es el número máximo de nodos que pueden ser padres de un nodo en la estructura de la red bayesiana en este caso el máximo número de padres que puede tener un nodo es de 4.

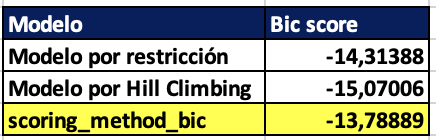
max\_iter: Es el número máximo de iteraciones que (HillClimb) realizará, en este caso int(1e4) es una notación abreviada para 10,000 iteraciones.

Red Bayesiana del modelo:



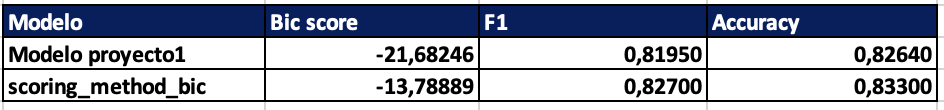
**Puntaje BIC modelo =** -13788.885951

Luego de tener los tres modelos se procedió a comparar los tres puntajes para encontrar el modelo con el BIC score menos negativo obteniendo la siguiente tabla y como mejor modelo el **scoring\_method\_BIC:**

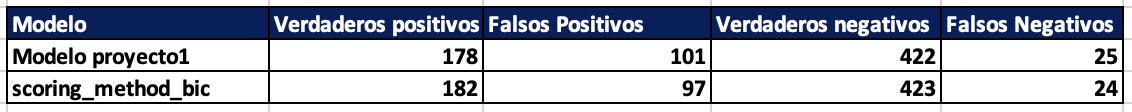


\*cifras en miles

1. Al evaluar el modelo del proyecto 1 y el modelo de seleccionado en el punto anterior obtuvimos por medio del código comparacion\_de\_modelos.ipynb la siguiente tabla con los siguientes valores:



\*cifras del BIC score en miles



**Métricas:**

1. **BIC Score:** Al ser un criterio que se emplea para comparar modelos estadísticos se tiene que para el modelo de bic scoring method se tiene un valor mucho menor por lo que se puede decir que este modelo se ajusta mejor a los datos.
2. **Accuracy o exactitud:** Esta métrica compara la proporción de predicciones correctas con el total de predicciones. Al tener un accuracy un poco mayor en el modelo de bic scoring method que sugiere que se tiene una predicción más precisa.
3. **Valor F1:** El Valor F1 combina el accuracy y la recall o sensibilidad en una sola métrica. El modelo bic scoring method al tener un valor ligeramente superior sobre el otro modelo sugiere que existe un mejor equilibrio entre precisión y recall.
4. **Verdaderos Positivos, Falsos Positivos, Verdaderos Negativos y Falsos Negativos:** El modelo bic scoring method se considera mejor ya que cuenta con más Verdaderos Positivos y menos Falsos Positivos y Falsos Negativos.

En Resumen, el modelo BIC scoring method en términos de todas las métricas nos muestra un modelo más confiable y preciso al momento de predecir y se ajusta mejor a los datos.