DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
Analítica Computacional para la Toma de Decisiones – Proyecto 2

PROFESOR: Juan F. Pérez

Apellidos	Nombres	Código	Login	Quién entrega (Bloque Neón)
Orjuela Lopez	Jeison Camilo	202224020	j.orjuelal	
Vega Medina	Jhon Gerardo	201418419	jg.vega10	✓

Link del Repositorio en Git Hub: https://github.com/JeisonCamiloO/DashPlotly2

Máquina virtual: **http://44.217.9.238:8050/**

Password BD RDS: 'proyecto' -> realizar el cambio directamente en el archivo **bd_conexion.py**

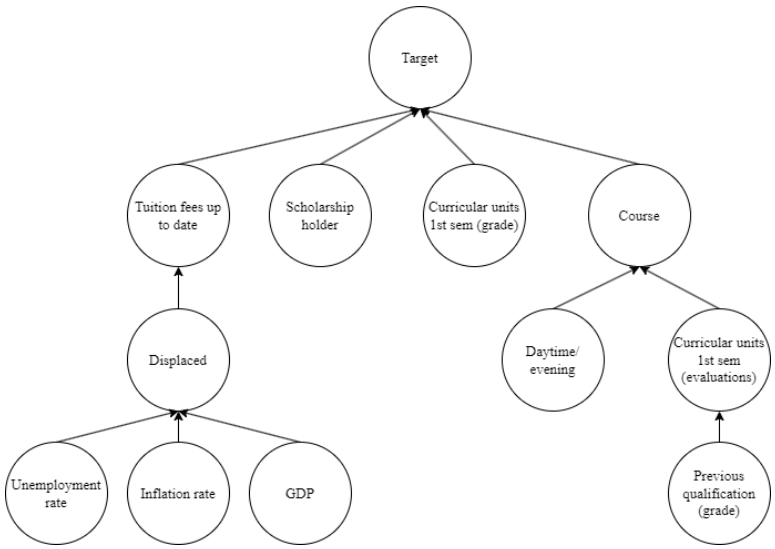
1. Modelos

- **Modelo original**

El modelo original hace referencia al modelo de predicción para el éxito académico realizado en la entrega del proyecto 1. Para su realización, se trabajó con la fuente de datos recolectada por la universidad de California en Irvine. Seguidamente, se escogieron aquellas variables que se consideran importantes, sustentadas por la literatura, y que impactaran la variable de interés **Target**, como también, su influencia en las decisiones del cliente: la universidad. Por último, se analizó la correlación entre todas las variables para identificar información redundante e identificar igualmente aquellas variables independientes que influyeran en mayor medida sobre la variable **Target**.

Luego de llevar a cabo el proceso de filtrado de todas las variables y, tras identificar las más relevantes, se procedió a construir la red Bayesiana. Los arcos entre los nodos se crearon gracias a su sustentación de la literatura y finalmente, el modelo se había construido. Por lo tanto, era necesario calcular métricas, como, por ejemplo: aciertos, tasa de aciertos, matriz de confusión, etc. Con estas medidas de medición se podía entender qué tan bueno había sido el modelo creado.

- **Red Bayesiana resultante**



- **Puntaje**

  El modelo original se valoró también por medio de la estimación del grafo de los puntajes k2 y BIC para poder hacerlo comparable entre los otros modelos realizados. A continuación, se presentan los puntajes resultantes para el modelo original.

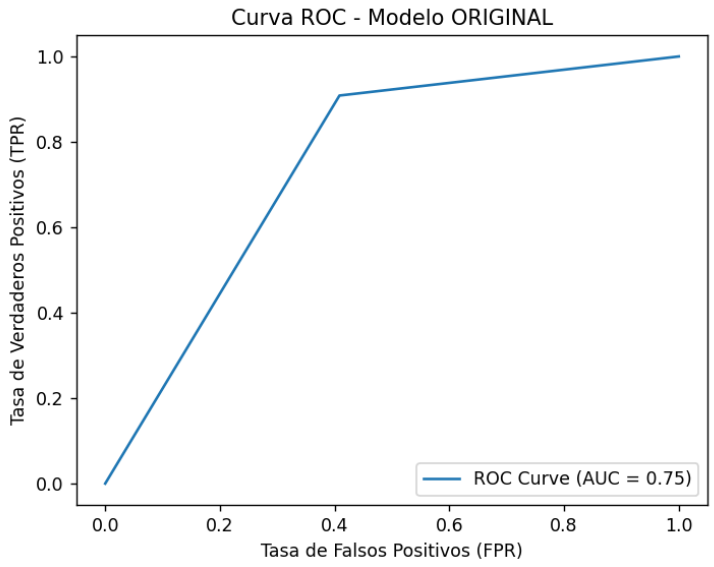
  - **K2:** -37084.24
  - **BIC:** -37955.5

- **Métricas**

  - **Aciertos:** 611
  - **Exactitud:** 69%
  - **Sensibilidad:** Dropout – 58.6%, Graduate – 90.85%, Enrolled – 19.1%
  - **Matriz de confusión:**

	Dropout	**Graduate**	**Enrolled**
Dropout	167	86	32
Graduate	19	417	23
Enrolled	26	88	27

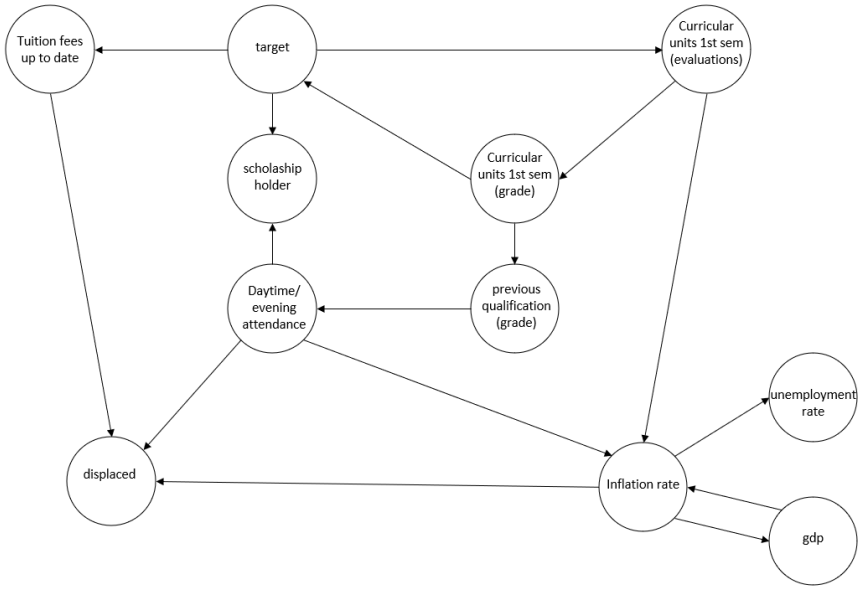
  - **Curva ROC:**



- **<u>Modelo por puntaje K2</u>**

Según las variables identificadas para realizar la red bayesiana del modelo original, se emplearon para estimar el grafo por medio del puntaje K2. Para lo anterior, se hizo uso de la librería **pgmpy** y se usaron las funciones ***HillClimbSearch()*** y ***K2Score()***. La estimación de las relaciones entre nodos se limitó a máximo 10 mil iteraciones, y 4 grados por nodo. Con esto, la idea es que el algoritmo identifique las relaciones entre los nodos que escoja y maximice el ajuste de los datos observados, lo que implica maximizar la probabilidad conjunta según la información suministrada.

- **Red Bayesiana resultante**



- **Puntaje**

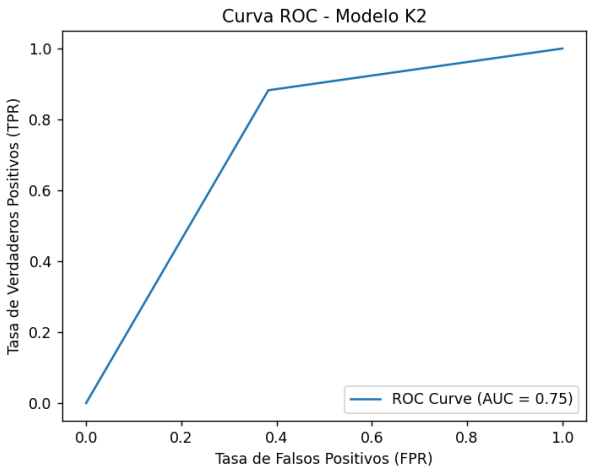
Este modelo construido por medio del puntaje K2 también se le calculó su puntaje BIC y los resultados son:

  - **K2:** -26.542
  - **BIC:** -26.315

- **Métricas**
  - **Aciertos:** 604
  - **Exactitud:** 68.2%
  - **Sensibilidad:** Dropout – 57.9%, Graduate – 88.2%, Enrolled – 24.1%
  - **Matriz de confusión:**

	Dropout	**Graduate**	**Enrolled**
Dropout	165	76	44
Graduate	18	405	36
Enrolled	20	87	34

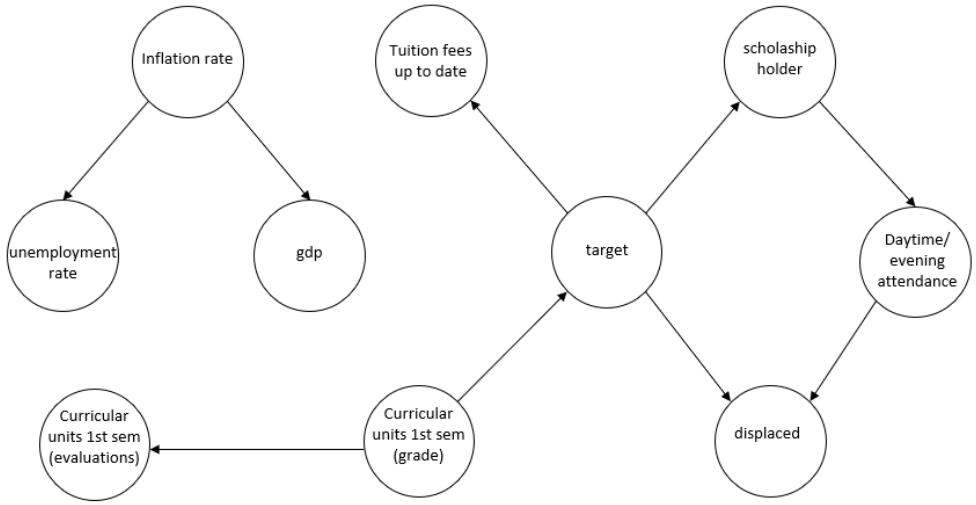
  - **Curva ROC:**



- **<u>Modelo por puntaje BIC</u>**

Para la realización de la estimación del grafo según las variables escogidas desde el modelo original se empleó el algoritmo de HillClimbSearch por medio del puntaje BIC. Para lo anterior, se hizo uso de las funciones ***HillClimbSearch()*** y ***BicScore()*** de la librería **pgmpy**. Al igual que con el puntaje K2, se limitó el algoritmo a 10 mil iteraciones y 4 grados por nodo. Con esto, se pretende que el software itere sobre la selección de arcos a agregar en el modelo para identificar aquellas conexiones entre variables que maximicen la probabilidad conjunta según la información suministrada.

- **Red Bayesiana resultante**



- **Puntaje**

Este modelo construido por medio del puntaje BIC también se le calculó su puntaje K2 y los resultados son:

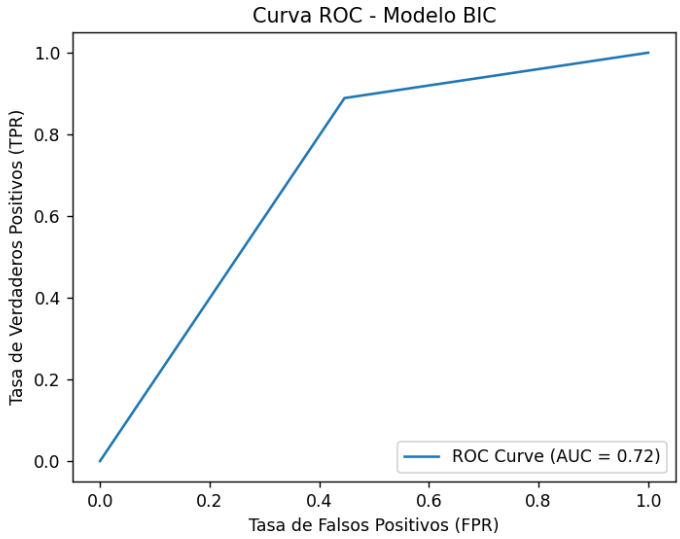
- **K2:** -24.890
- **BIC:** -26.710

- **Métricas**

- **Aciertos:** 590
- **Exactitud:** 66.6%
- **Sensibilidad:** Dropout – 62.8%, Graduate – 88.9%, Enrolled – 2.1%
- ***Matriz de confusión:***

	Dropout	**Graduate**	**Enrolled**
Dropout	179	95	11
Graduate	43	408	8
Enrolled	43	95	3

- **Curva ROC:**



2. Comparación modelos

Después de haber construido los 3 modelos y de calcular sus métricas es necesario realizar una comparación entre ellos para identificar el mejor modelo con el cual se podrá trabajar y realizar mejores estimaciones.

Para el modelo original se calculó su respectivo puntaje K2 y BIC para que fuera comparable con los modelos estimados por medio de estos 2 puntajes. Entonces, para el modelo original se obtuvo un score K2 de -37084.24 y el para modelo con puntaje K2 se obtuvo un score igual a -26542.06. Por lo tanto, se evidencia que por puntaje el mejor modelo es el K2. Por otro lado, para realizar la comparación con el modelo original y el modelo estimado por el puntaje BIC se obtuvo un score de -37955.5 y -26710.33 respectivamente. Por lo tanto, el mejor modelo es el estimado por puntaje BIC ya que es el valor más cercano a cero.

Sin embargo, para cada uno de los modelos también se calcularon métricas como: exactitud, aciertos, curva ROC y sensibilidad que también ayudarán a escoger aquel modelo que sea mejor sobre los otros.

				Sensibilidad		
Modelo	**Puntaje BIC**	**Puntaje K2**	**Exactitud**	**Dropout**	**Enrolled**	**Graduated**
Original	-37.955	-37.084	**69%**	59%	19%	91%
BIC	-26.710	-24.890	**67%**	63%	2%	89%
K2	-26.315	-26.542	**68%**	58%	24%	88%

En este caso, se decidió que fuera la **exactitud** la métrica diferenciará el mejor modelo, por ende, un modelo con mayor exactitud se preferirá. Así las cosas, para el modelo original, K2 y BIC se obtuvo una exactitud de 69%, 68.2% y 66.6% respectivamente. De manera que el modelo que sobresale sobre los otros y que tiene una mayor capacidad predictiva sobre el éxito estudiantil será el **modelo original**.

3. Conclusión

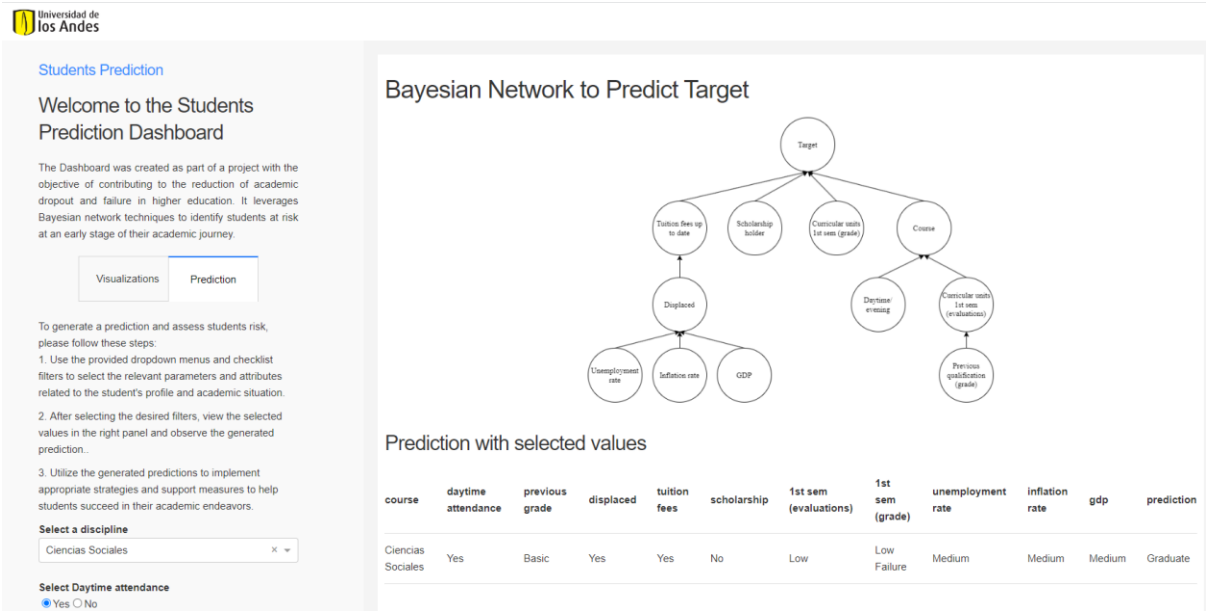
A pesar de que los 3 modelos construidos tienen valores para sus métricas bastante parecidos se decidió por el de mayor exactitud. Sin embargo, si nos fijamos en la sensibilidad de cada uno de los modelos, por ejemplo, para el modelo K2 se identifica una mayor sensibilidad en la categoría Enrolled que implica una mayor asertividad en la predicción de este nivel de la variable de interés, pero no clasifica mejor ni Dropout ni Graduated. Dado que las categotias

mas importante según el cliente, las universidades, se da mayor relevancia a los niveles Dropout y Graduated. Por lo anterior, Se escogerían los modelos BIC y el original y, entre ellos dos, el que mejor desempeño tiene en predicción es el original, por ende, el producto para la predicción dirigido a las universidades se realiza sobre este modelo original.

Anexo – Producto final

Utilizando el mejor modelo se desarrolló una aplicación web para universidades que les permita predecir el éxito académico de sus estudiantes. Esta herramienta se encuentra organizada por dos pestañas en las que se puede encontrar algunas visualizaciones que van a permitirle al usuario final hacer un diagnóstico inicial y una predicción que de acuerdo con las características de un estudiante estima una predicción y determina si este estudiante se graduará, deserta o continuará matriculado al haber transcurrido los cuatro años de formación académica.

Pestaña de visualizaciones



Pestaña de predicción