María Gabriela Urango Llorente - 202013781
Andrea Velandia León - 201817500

Analítica Computacional para la Toma de Decisiones

Proyecto 1

1. Delimitación del problema

Según el Ministerio de Educación Nacional, la tasa de deserción anual universitaria en Colombia es del 8,02% (EL TIEMPO, 2023), la cual ha ido en aumento en los últimos años. Lo anterior es una cifra alarmante ya que el abandono de la educación superior afecta negativamente tanto el futuro de los estudiantes como del desarrollo del país (Pontificia Universidad Javeriana, 2023). Las causas de la deserción escolar vienen relacionadas principalmente al reto de adaptación de los estudiantes a la vida universitaria. Donde, la exigencia y autonomía del aprendizaje, así como la convivencia en el entorno universitario tiene un impacto en su rendimiento académico, llevándolos a la deserción escolar (Martins et al., 2023).

Por lo anterior, es importante, para las instituciones de educación superior, el diseño y ejecución de estrategias que permitan un monitoreo constante y sistemas de detección tempranas para identificar las causas de la deserción, y así, evitarla de ser posible. Esto puede incluir programas de adaptación para nuevos estudiantes, asesoramiento académico, consejería estudiantil, servicios de tutorías, entre otros.

Por lo tanto, el presente proyecto se enfoca en el desarrollo de una herramienta para la toma de decisiones basada en datos enfocada en las universidades que permita la detección temprana de la deserción escolar. Para que así, las universidades puedan desarrollar políticas de ayuda para disminuir la problemática, y de esta manera, evitar la pérdida de recursos económicos que incluye la deserción. Por ello, se busca responder las siguientes preguntas de negocio:

- ¿Qué factores socio demográficos, académicos y financieros influyen en la deserción académica de los estudiantes universitarios?
- ¿Cuál es la probabilidad de que un estudiante abandone su programa académico en los primeros semestres?

2. Objetivos

a. Objetivo general:

Desarrollar una herramienta de analítica de datos para la toma de decisiones que permita la detección temprana de la deserción estudiantil para que las instituciones universitarias puedan invertir en estrategias enfocadas en el éxito académico de los estudiantes.

b. Objetivos específicos:

i. Analizar la información histórica del comportamiento estudiantil recolectada por la universidad en la base de datos *Predict students' dropout and academic success*.
ii. Proponer un modelo de red Bayesiana que permita inferir la probabilidad de deserción de un estudiante.

iii. Implementar el modelo predictivo en una interfaz interactiva para la evaluación de distintos casos de estudiantes universitarios.

3. Estadísticas descriptivas

La información con la que se cuenta es una base de datos *Predict students' dropout and academic success,* el cual es un conjunto de datos creado a partir de una institución de educación superior que contiene información de estudiantes matriculados en diferentes carreras universitarias. El conjunto de datos incluye tanto información en el momento de la inscripción de los estudiantes (ruta académica, demografía y factores socioeconómicos) como el rendimiento académico de los estudiantes al final del primer y segundo semestre.

Por un lado, la base de datos cuenta con 4424 observaciones de estudiantes, de los cuales 2209 se graduaron, 1421 se retiraron y 794 estudiantes siguen matriculados.

Ilustración 1-Histograma de la variable de interés: target

Por otro lado, para evidenciar posibles relaciones causales entre las variables del dataset y la variable de interés (Target: vinculo actual del estudiante con la universidad: graduado, retirado o matriculado) se realiza la siguiente tabla de correlación.

María Gabriela Urango Llorente - 202013781
Andrea Velandia León - 201817500

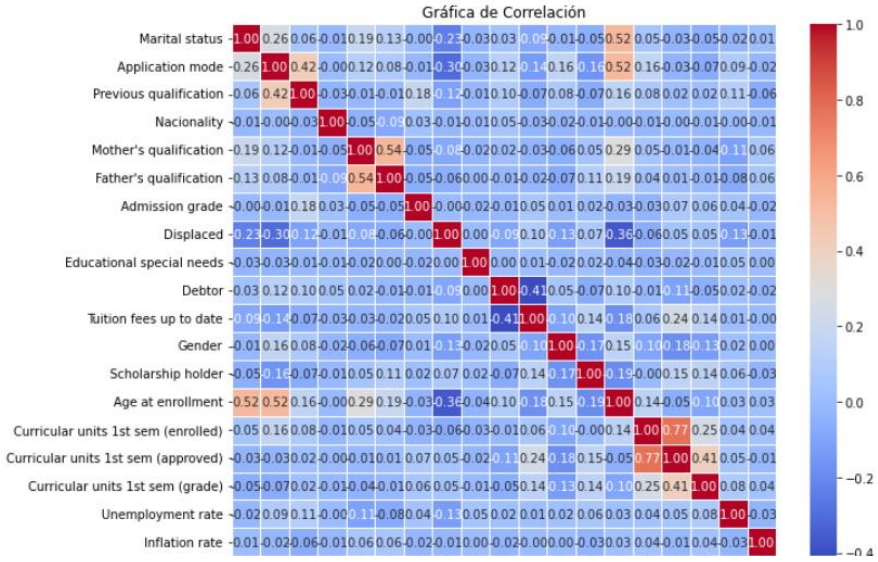
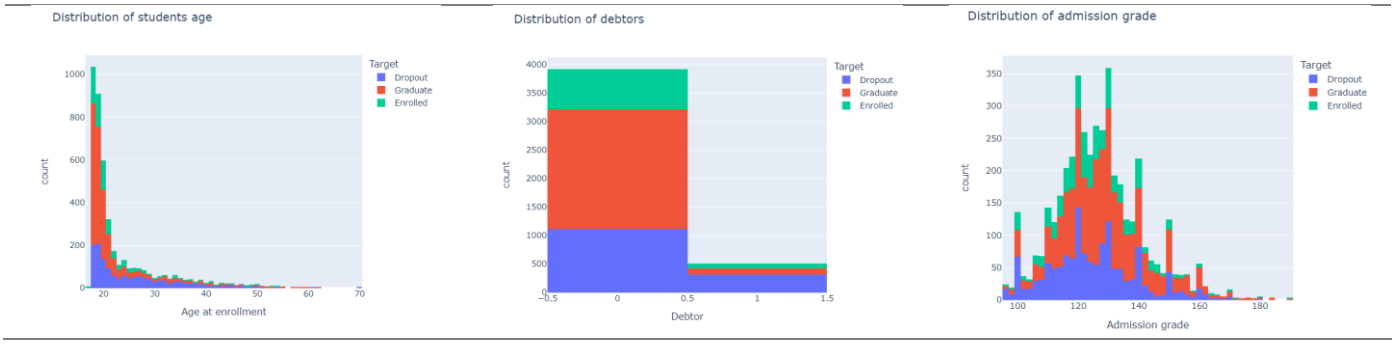


Ilustración 2-Matriz de correlación para todas las variables del modelo

Y a pesar de que se está correlacionando también variables discretas, se evidencia que hay relación entre:

- Las variables que reflejan los resultados del primer semestre académico del estudiante.
- El modo de aplicación con la edad en la que se matricula el estudiante y sus estudios previos.
- La edad en la que se matricula el estudiante y el estado civil.
- Las variables financieras como si el estudiante está vinculado a un crédito financiero y si está a paz y salvo para pagar la universidad.
- Entre otras correlaciones de menor significancia.

Por otro lado, se analiza la relación entre las distintas variables (que intuitivamente se asocian por ser determinante en la deserción universitaria) con la variable *Target.* Observamos que las siguientes 9 variables tienen una tendencia a tener una influencia en la variable de acuerdo con los valores que adopte:



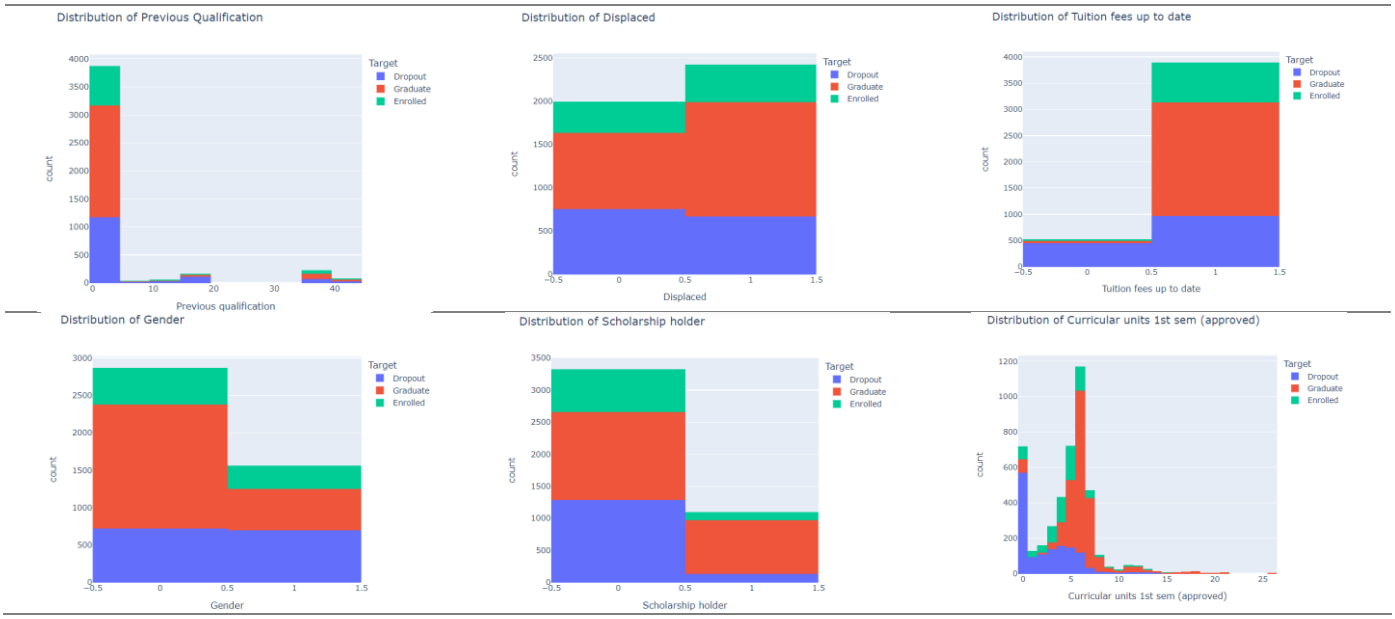


Tabla 1 -Histogramas de las variables relacionadas frente a Target

Observamos que en las otras variables no se evidencia una relación o no hay suficientes datos para realizar una inferencia:

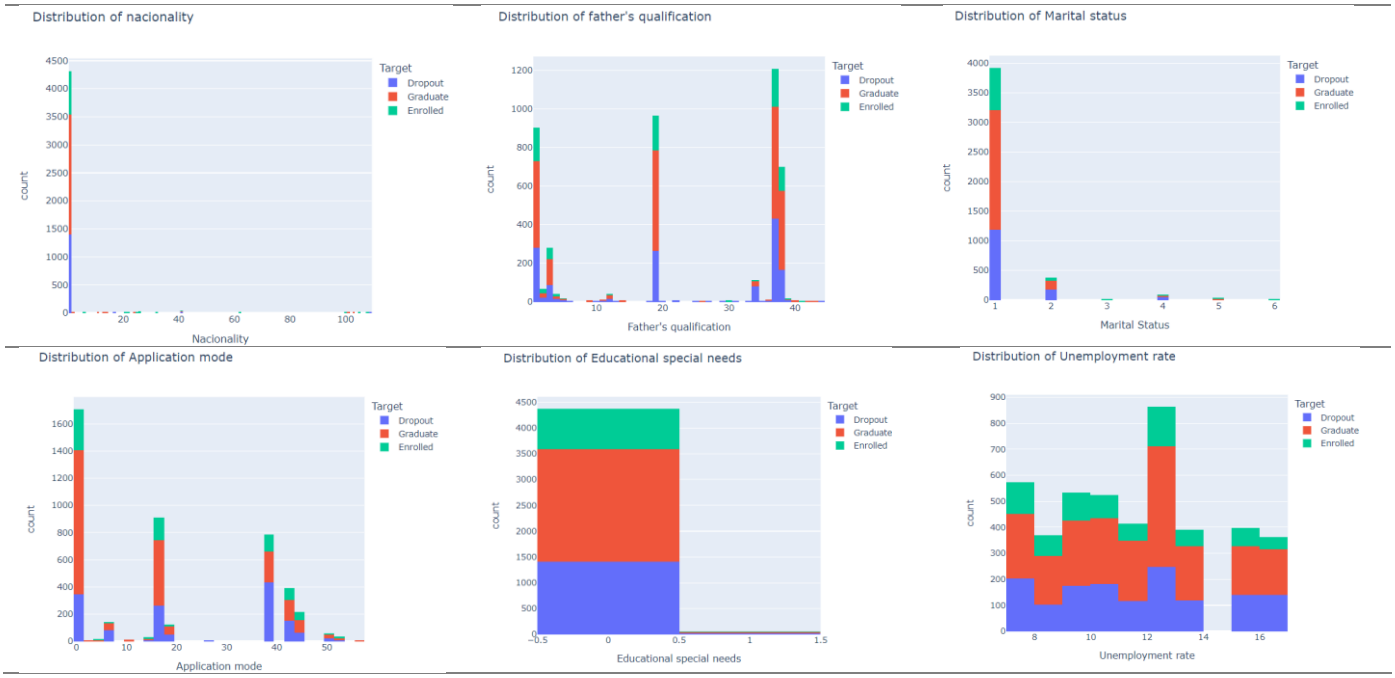


Tabla 2 -Histogramas de las variables descartadas frente a Target

4. Modelo

a. Selección de variables

i. Revisión bibliográfica

Las causas del fracaso y deserción académica en las instituciones de educación superior se han estudiado en todo el mundo e incluyen tanto dificultades académicas como factores financieros, personales y familiares (Martins et al., 2023).

Algunos artículos se basan únicamente en variables académicas que afectan la deserción como el porcentaje de asistencias a clase, las notas de las calificaciones finales de los cursos, el promedio GPA del estudiante, el número total de créditos aprobados, el área escolar, el semestre actual cursado, el programa en el que se encuentre matriculado, entre otros (Beaulac & Rosenthal, 2019; Fernandez-Garcia et al., 2021; Hutagaol & Suharjito, 2019). Otros estudios agregan factores sociodemográficos y socioeconómicos que afectan la deserción como el sexo del estudiante, la edad, los ingresos y la educación de los padres, la situación laboral actual, el lugar de nacimiento, entre otras (Dien et al., 2020; Hutagaol & Suharjito, 2019).

ii. Variables seleccionadas:

Teniendo en cuenta tanto la base de datos y las estadísticas descriptivas analizadas como la revisión bibliográfica de los factores influyentes en la deserción de los estudiantes, es pertinente tener en cuenta en un modelo predictivo tanto variables en una dimensión socio demográfica, financiera como académica. Las variables seleccionadas y de interés a evaluar en cada dimensión son las siguientes:

- **Dimensión socio demográfica:**
  - *Age at enrollment*: Edad en que el estudiante se matriculó a la universidad.
  - *Gender:* Genero del estudiante.
  - *Displaced:* 1 si el estudiante proviene de otro lugar diferente a la ubicación de la universidad, 0 d.l.c.

- **Dimensión financiera y económica:**
  - *Scholarship holder*: 1 si el estudiante es becado en la universidad, 0 d.l.c.
  - *Debtor:* 1 si el estudiante está vinculado a un crédito financiero para pagar la universidad, 0 d.l.c.
  - *Tuition fees up to date:* 1 si el estudiante esta al día con el pago de matrícula de la universidad, 0 d.l.c.

- **Dimensión académica:**
  - *Admission grade*: Nota de admisión para entrar a la universidad.
  - *Curricular units 1st sem (grade):* Promedio global de los cursos vistos en el primer semestre de la universidad.
  - *Curricular units 1st sem (approved):* Número de cursos aprobados en primer semestre de la universidad.

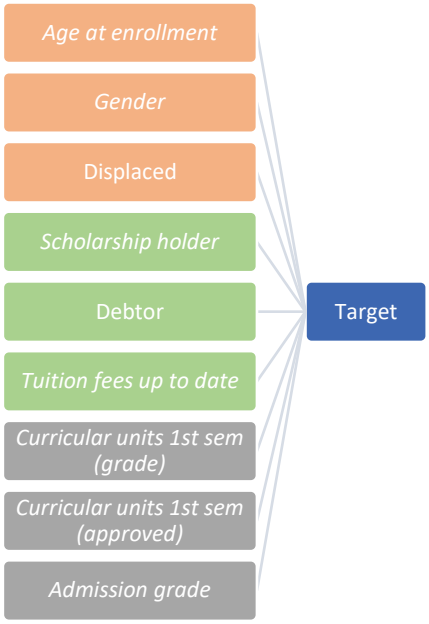
Por otro lado, la variable de interés es *Target,* la cual indica si el estudiante de interés se graduó o desertó de la universidad.

b. Selección del modelo

Para la predicción de la deserción de los estudiantes (Target), se implementa un modelo de red Bayesiana el cual permite determinar las posibles relaciones causales (dependencia) entre las variables seleccionadas y la variable de interés.

Modelo 1:

Un primer modelo planteado tiene una estructura de Naive Bayes en el cual consideramos todas las variables anteriormente seleccionadas en las dimensiones sociodemográficas, económicas y académicas para predecir la deserción escolar. En el modelo consideramos la nota y los créditos aprobados a lo largo del primer año académico porque el objetivo es desarrollar un sistema que ayude a segmentar a los estudiantes lo antes posible desde el inicio de su trayectoria en la educación superior.

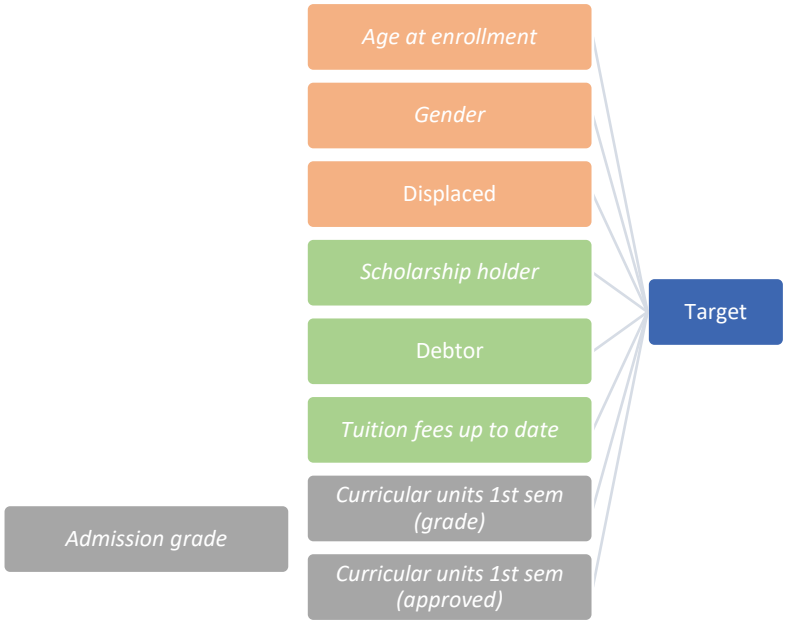


Estructura del modelo 1 de red Bayesiana para la predicción de la deserción universitaria

Modelo 2:

Un aspecto para considerar en el anterior modelo es que es necesario contar con la información de los resultados finales del estudiante en su primer semestre de la universidad. Lo anterior podría ser un aspecto limitante a la hora de analizar y proponer estrategias de ayuda a los estudiantes que están a punto de ingresar a la universidad o están en sus primeras semanas (primíparos y momento donde la adaptación a la vida universitaria es fundamental). Por lo tanto, se propone una estructura de red Bayesiana con cambios en las variables académicas. Donde, la nota de admisión (que se conoce a priori) tiene una influencia a la nota y créditos aprobados en primer semestre, y estos a la vez tienen una relación con la deserción universitaria. De esta manera, la universidad podría predecir tanto los resultados

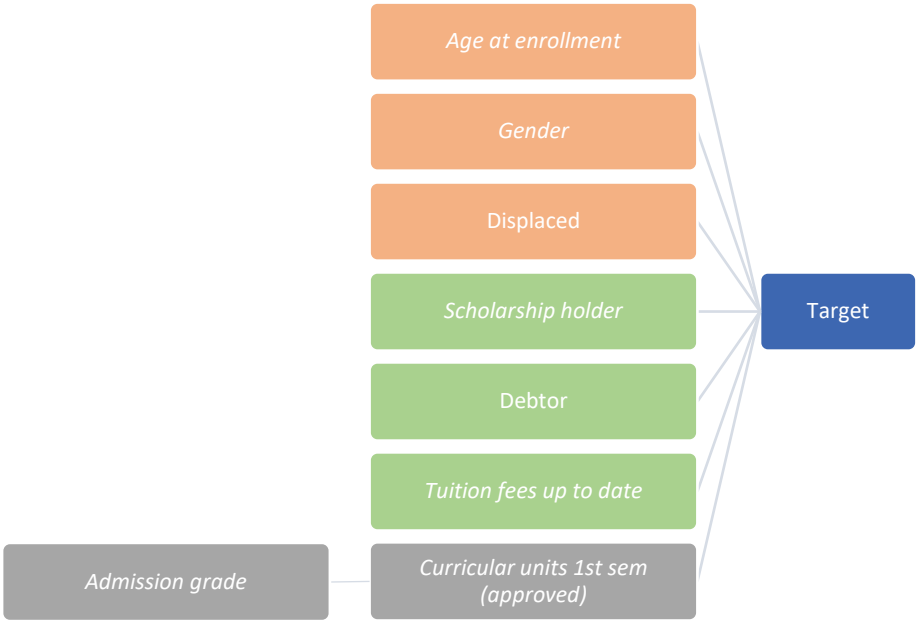
del primer semestre como la probabilidad de deserción con solo contar con la nota de admisión a la universidad.



Estructura del modelo 2 de red Bayesiana para la predicción de la deserción universitaria

Modelo Final:

Ahora bien, analizando la tabla de correlaciones entre variables, se evidencia que hay una considerable correlación positiva entre las variables académicas de los resultados del primer semestre: la nota final del primer semestre y los créditos aprobados. Por lo tanto, para alimentar el supuesto de independencia entre las variables, se decidió eliminar una de las dos variables. Por lo tanto, el modelo final de Red Bayesiana para la predicción de la deserción universitaria es el siguiente:



. Estructura del modelo fila de red Bayesiana para la predicción de la deserción universitaria

c. Transformación de los datos

Es pertinente mencionar la transformación de algunas variables realizada en la base de datos. Por un lado, teniendo en cuenta que la variable de interés (target) no solo considera los valores de personas graduadas y retiradas sino también los que actualmente estan matriculados, se deicidio eliminar las observaciones en que el estudiante aún está matriculado pues no es de interés para la pregunta seleccionada.

Por otro lado, a algunas variables seleccionada fue necesario categorizarlas en otras variables para 1.) evitar variables con demasiados valores y 2.) discretizar variables numéricas. Las variables que se transformaron en diferentes rangos fueron: *Age at enrollmente, Curricular units 1st sem (approved), Curricular units 1st sem (grade)* y *Admission grade*. Para discretizar las variables fue de ayuda los histogramas pues permitieron identificar comportamientos similares entre diferentes valores de las variables.

d. Estimación de parámetros

Para poder hallar la estimación de los parámetros de la Red Bayesiana (distribuciones de probabilidad y distribuciones de probabilidad condicionales) de cada variable se empleó el método de **estimación por máxima verosimilitud**. Para ello, previamente los datos fueron clasificados en datos de prueba y dados de entrenamiento del modelo con un 20% y 80% respectivamente.

Algunos de los resultados de la estimación de los parámetros son los siguientes:

Gender(0)	0.651171
Gender(1)	0.348829

Debtor	...	Debtor(1)
Displaced	...	Displaced(1)
Gender	...	Gender(1)
Scholarship holder	...	Scholarship holder(1)
Tuition fees up to date	...	Tuition fees up to date(1)
age_range	...	age_range(Rango >41)
approved_sem1_range	...	approved_sem1_range(Rango >10)
Target(Dropout)	...	0.5
Target(Graduate)	...	0.5

Displaced(0)	0.450758
Displaced(1)	0.549242

Scholarship holder(0)	0.735537
Scholarship holder(1)	0.264463

Debtor(0)	0.884642
Debtor(1)	0.115358

Tuition fees up to date(0)	0.134298
Tuition fees up to date(1)	0.865702

age_range(Rango 16-18)	0.237603
age_range(Rango 19-20)	0.334366
age_range(Rango 21-23)	0.134986
age_range(Rango 24-30)	0.134298
age_range(Rango 31-40)	0.102617
age_range(Rango >41)	0.0561295

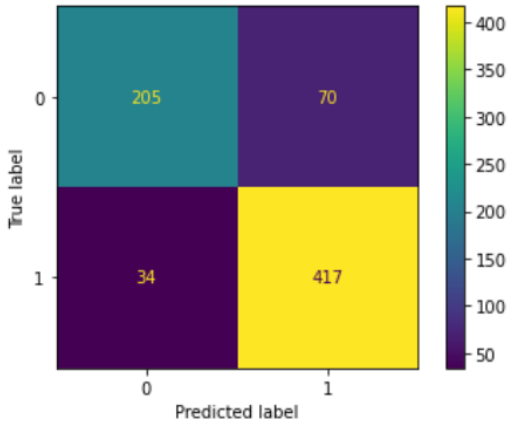
grade_admission_range	...	grade_admission_range(Rango >160)
approved_sem1_range(Rango 0-2)	...	0.25
approved_sem1_range(Rango 3-5)	...	0.22826086956521738
approved_sem1_range(Rango 6-10)	...	0.4673913043478261
approved_sem1_range(Rango >10)	...	0.05434782608695652

grade_admission_range(Rango 0-100)	0.00998623
grade_admission_range(Rango 100-115)	0.165634
grade_admission_range(Rango 115-130)	0.42803
grade_admission_range(Rango 130-145)	0.277893
grade_admission_range(Rango 145-160)	0.0867769
grade_admission_range(Rango >160)	0.0316804

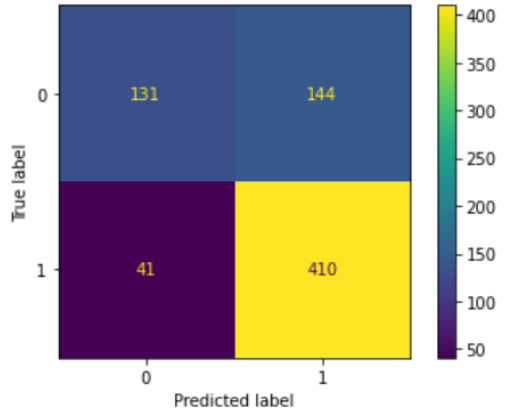
5. Predicción

Finalmente, la precisión de predicción del modelo seleccionado fue evaluada a través de los datos de prueba.

Por un lado, al evaluar el modelo con **todas** las variables, se obtiene una predicción con 487 estudiantes que se graduaron y 239 estudiantes que se retiraron de la universidad. También, un *accuracy* y *F1 score* de 0.856 en la exactitud de la predicción. Asimismo, se obtiene la siguiente matriz de confusión:



Por otro lado, al evaluar el modelo sin la variable con padre (Curricular units 1st sem (approved)), se obtiene una predicción con 554 estudiantes que se graduaron y 172 estudiantes que se retiraron de la universidad. Tambien, un *accurency* y *F1 score* de 0.745 en la exactitud de la predicción. Asimismo, se obtiene la siguiente matriz de confusión:



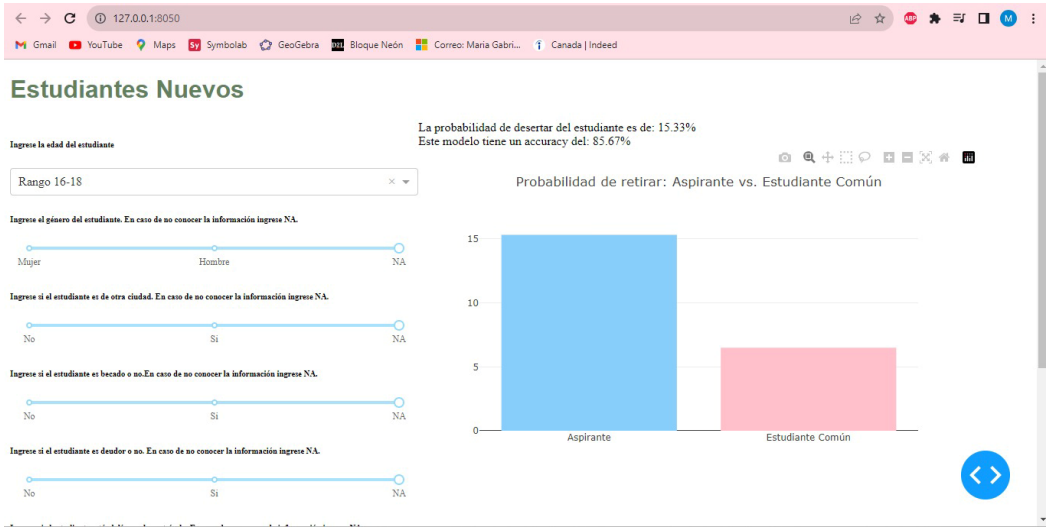
Sin esta información de los resultados del primer semestre se evidencia que la precisión de la predicción disminuye un poco, donde principalmente aumentan los casos de falsos positivos.

6. Herramienta de toma de decisiones

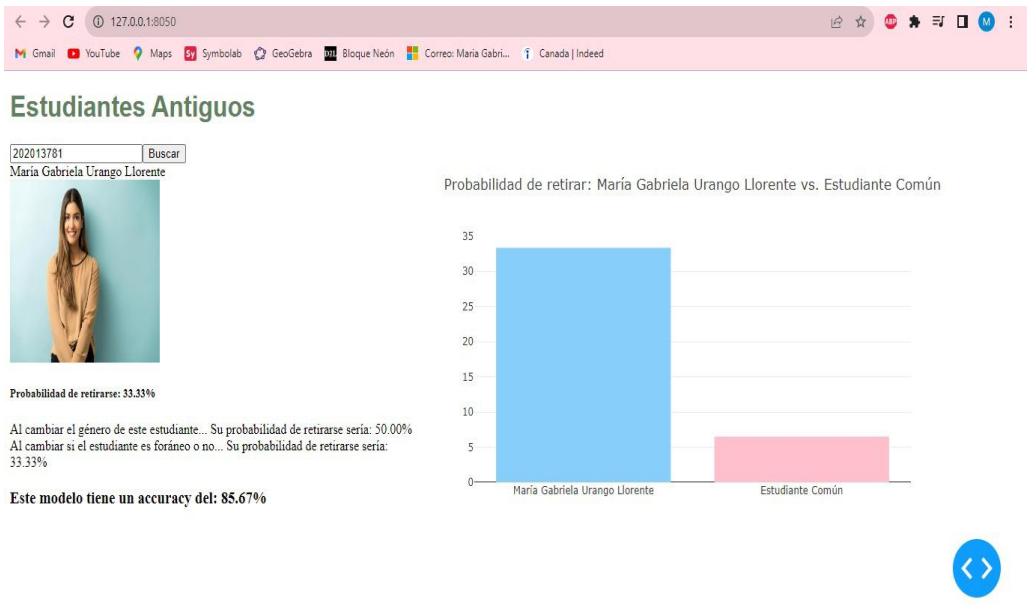
Se espera que el anterior modelo de predicción de la deserción universitaria pueda ser utilizado para la toma de decisiones en las instituciones universitarias. Por lo tanto, el modelo fue integrado en una interfaz/monitor para que la universidad pueda usarlo como herramienta para predecir la probabilidad de deserción de cualquier estudiante.

La herramienta cuenta con dos interfaces, la primera para estudiantes nuevos y la segunda para estudiantes antiguos y registrados en el sistema de información de la universidad.

- La primera interfaz permite que la universidad ingrese manualmente todos los datos de interés del estudiante (rango de edad en que entrará a la institución, género, variables de las condiciones financieras, si el estudiante es de otra ciudad y la nota de admisión). La herramienta retorna tanto la probabilidad de desertar del estudiante, el accuracy del modelo y una comparación de la probabilidad de deserción entre el estudiante y el estudiante promedio de la universidad.



- La segunda interfaz permite que la universidad únicamente tenga que ingresar el código de identificación del estudiante. Debido a que, la universidad ya debería tener guardada toda la información del estudiante y la herramienta se alimentaría de ese sistema de información. La herramienta retorna la información general y la fotografía del estudiante, la probabilidad de deserción, el accuracy del modelo, la comparación entre el estudiante seleccionado y el estudiante promedio, así como algunos insights



de un análisis de sensibilidad entre las variables seleccionadas y la probabilidad de deserción.

7. Recomendaciones y conclusiones

- Tanto los factores de las dimensiones demográficos, sociales, financieros y académicos influyen en la deserción de los estudiantes universitarios. Las principales variables que afectan la deserción son: la edad en que se matricula la persona, el género, si el estudiante procede de otra ciudad, si es deudor, si está a paz y salvo de las obligaciones financieras, si es becado y el número de número de créditos aprobados en primer semestre por el estudiante.

- Un modelo de Red Bayesiana permite predecir la deserción universitaria de los estudiantes desde el inicio de su trayectoria universitaria. El modelo con todas las variables (medido una vez finalizado el primer semestre del estudiante) tiene una eficacia de la predicción de 85.6%. No obstante, la eficacia de la predicción disminuye aproximadamente un 10% si no se tienen los resultados académicos del primer semestre del estudiante.

- Se recomienda a la universidad emplear el modelo de predicción de deserción en cada uno de sus estudiantes. Puesto que, esté le permitirá una detección temprana de estos casos, y de esta manera, disminuir los riesgos tanto económicos como de reputación al poder emplear diferentes estrategias de acompañamiento, ayuda y consegeria a los estudiantes.

8. Repositorio en GitHub:

https://github.com/mgabyurango/Proyecto1-ACTD

9. Reporte de trabajo en equipo

Para el desarrollo del proyecto la siguiente hoja de ruta con su respectiva distribución de tareas fue establecida entre las integrantes del equipo (Gabriela y Andrea):

- Delimitación del problema y definición de la pregunta de interés (Gaby y Andre).
- Análisis descriptivo (Gaby y Andre).
- Modelación e inferencia (Andre)
- Herramienta en Dash (Gaby).
- Informe y presentación final (Gaby y Andre).

No obstante, finalmente todos los integrantes terminaron contribuyendo en todas las etapas del desarrollo.

10. Bibliografía

Beaulac, C., & Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, *60*(7). https://doi.org/10.1007/s11162-019-09546-y

Dien, T. T., Luu, S. H., Thanh-Hai, N., & Thai-Nghe, N. (2020). Deep learning with data transformation and factor analysis for student performance prediction. *International Journal of Advanced Computer Science and Applications*, *11*(8). https://doi.org/10.14569/IJACSA.2020.0110886

EL TIEMPO. (2023, June 26). *Tristes estadísticas de deserción universitaria: mitad de estudiantes no se gradúa*. Https://Www.Eltiempo.Com/Vida/Educacion/Tristes-Estadisticas-de-Desercion-Universitaria-Mitad-de-Estudiantes-No-Se-Gradua-789914.

Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3115851

Hutagaol, N., & Suharjito. (2019). Predictive modelling of student dropout using ensemble classifier method in higher education. *Advances in Science, Technology and Engineering Systems*, *4*(4). https://doi.org/10.25046/aj040425

Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education. *Applied Sciences*, *13*(8), 4702. https://doi.org/10.3390/app13084702

Pontificia Universidad Javeriana. (2023). *Panorama de la deserción en educación superior*.