**Predictor académico – Reporte**

*Daniela Andrea Ruiz López - 201914337*

*Juan Daniel Umaña Caro - 201315747*

**Pregunta de negocio:**

¿Cuáles variables pueden considerarse factores de riesgo para que un estudiante no tenga éxito académico en la universidad y cómo predecir la necesidad de acompañamiento de un estudiante a partir ellos?

**Metodología**

El predictor académico ha sido desarrollado con datos recolectados por el estudio Early prediction of student's performance in higher education: a case study.

En su primera versión el predictor utilizó las variables continuas y categóricas con mayor correlación sobre la variable objetivo: haberse graduado como estado de éxito académico.

Posteriormente estos predictores son utilizados en una red bayesiana para parametrizar un modelo predictivo el cual sea capaz de inferir el posible desempeño que tendrá un estudiante en su carrera a partir de su información actual.

Por último, el producto se ofrece como un tablero interactivo alojado en un servidor AWS para que usuarios puedan acceder al modelo de forma remota y evaluar las predicciones para sus estudiantes.

**Resultados**

*Selección de variables*

El estudio contempla 36 variables con potencial predictor y una variable target categórica que denota si el estudiante se gradúo (éxito), se retiró o aún se encuentra matriculado (fracaso). Las variables se pueden agrupar de la siguiente forma:

*Variables contínuas:*

* application\_order\_dic = Application order (between 0 - first choice; and 9 last choice)
* prev\_qual\_grade\_dic = Grade of previous qualification (between 0 and 200)
* admission\_grade = Admission grade (between 0 and 200)
* age = Age of studend at enrollment
* units\_i\_semester\_credited/enrolled/evaluations/approved/wo\_evaluation = number of units redited/enrolled/evaluations/approved/wo\_evaluation in semester i
* units\_i\_semester\_grade = Grade average in the ith semester (between 0 and 20)
* unemployment\_rate
* inflation\_rate
* GDP

*Variables categóricas:*

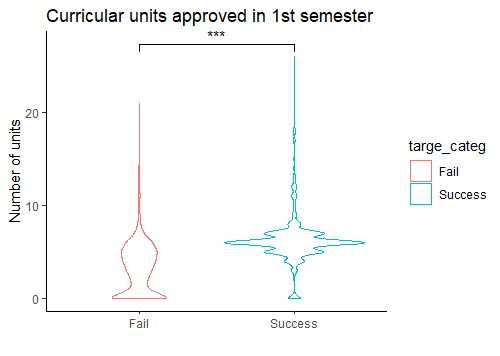
* 'marital\_status'
* 'application\_mode'
* 'application\_order'
* 'course'
* 'daytime/evening\_attendance'
* 'previous\_qualification'
* 'nacionality'
* 'mothers\_qualification'
* 'fathers\_qualification'
* 'mothers\_occupation'
* 'fathers\_occupation'
* 'displaced'
* 'educational\_special\_needs'
* 'debtor'
* 'tuition\_fees\_up\_to\_date'
* 'gender'
* 'scholarship\_holder'
* 'age\_at\_enrollment'
* 'international'

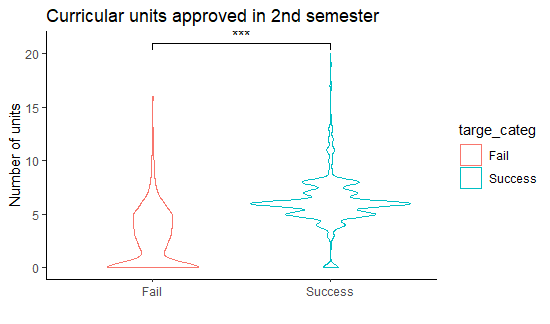
De acuerdo con las correlaciones en las variables continuas se decidió tomar las siguientes variables para las siguientes etapas del modelo:

* units\_i\_semester\_approved (positive correlation)
* units\_i\_semester\_grade (positive correlation)
* course (positive correlation)
* tuition\_fees\_up\_to\_date (positive correlation)
* scholarship\_holder (positive correlation)
* application\_mode (positive correlation)
* age\_at\_enrollment (negative correlation)
* units\_i\_semester\_wo\_evaluation (negative correlation)
* inflation\_rate (negative correlation)

A continuación, gráficos y análisis descriptivos de los predictores seleccionados:

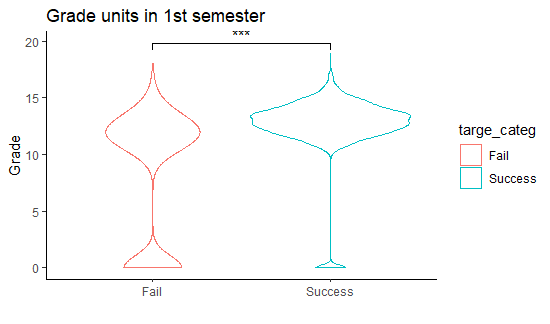
*Unidades curriculares aprobadas en el primer y segundo semestre*:

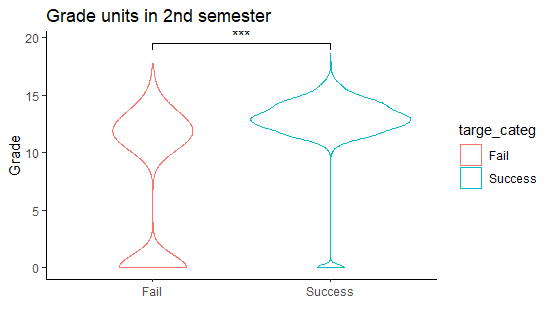




En ambos semestres las distribuciones son similares a través de ambos grupos, y se observa que existe una diferencia significativa entre el (mayor) número de unidades aprobadas por aquellos estudiantes que lograron éxito académico.

*Calificaciones de las unidades aprobadas en el primer y segundo semestre*





En ambos semestres las distribuciones son similares a través de ambos grupos, y se observa que existe una diferencia significativa entre las calificaciones de estudiantes que lograron éxito académico.

*Cursos*

Los cursos más frecuentes (y su porcentaje respecto a toda la población) de los estudiantes con éxito académico fueron:

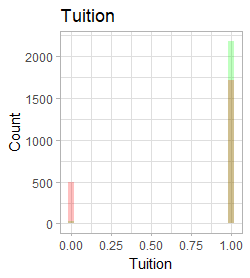
* Nursing (12.4%)
* Social Service (5.6%)
* Journalism and Communication (4.4%)

Los cursos más frecuentes (y su porcentaje respecto a toda la población) de los estudiantes sin éxito académico fueron:

* Management (5.5%)
* Nursing (4.9%)
* Management (evening attendance) (4.3%)

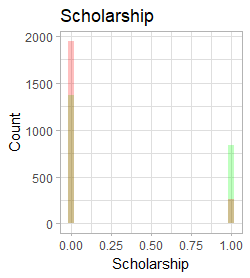
Nota: Para las siguientes gráficas de barras, la categoría de éxito (graduación) está representada con el color verde y la categoría contraria (estudiantes que desertaron o siguen matriculados) con el color rojo.

*Matricula*



Al agrupar los datos por rendimiento académico se observa que la mayoría de estudiantes que no se encuentran al día con su matrícula fueron estudiantes que no se graduaron.

*Beca*



Al agrupar los datos por rendimiento académico se observa que la mayoría de estudiantes que poseen una beca lograron éxito académico.

*Tipo de aplicación*

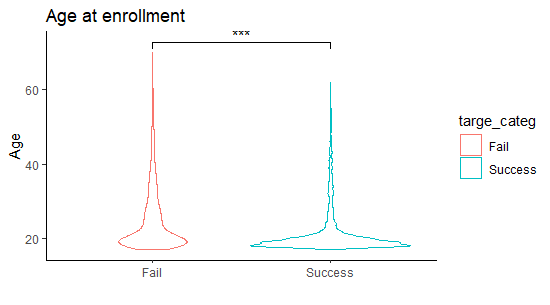
Los tipos de aplicación más frecuentes (y su porcentaje respecto a toda la población) de los estudiantes con éxito académico fueron:

* 1st phase: general contingent (24.0%)
* 2nd phase: general contingent (10.3%)
* Over 23 years old (5.2%)

Los tipos de aplicación más frecuentes (y su porcentaje respecto a toda la población) de los estudiantes sin éxito académico fueron:

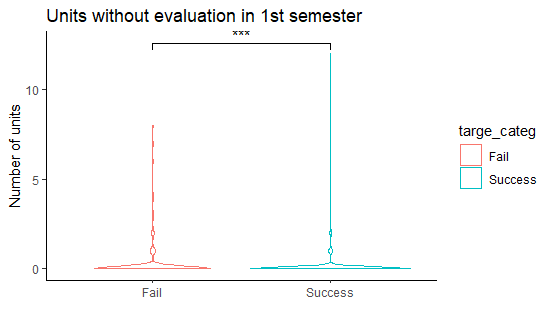
* 1st phase: general contingent (14.6%)
* Over 23 years old (12.6%)
* 2nd phase: general contingent (9.4%)

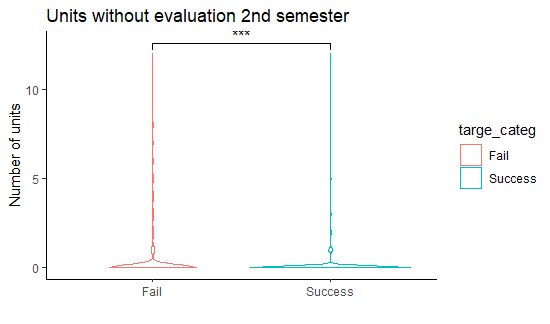
*Edad al momento de inscripción*



Existe una diferencia significativa entre las edades de los estudiantes con distinto desenlace académico, siendo el grupo de estudiantes que logró éxito académico aquel con menor edad.

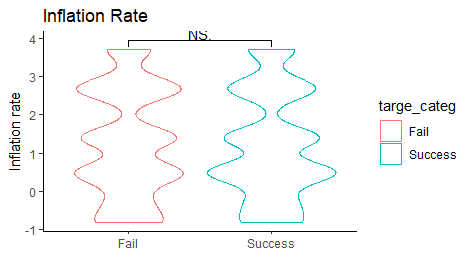
*Unidades sin evaluación en el primer y segundo semestre*





En ambos semestres las distribuciones son similares a través de ambos grupos, y se observa que existe una diferencia significativa entre el número de unidades aprobadas por aquellos estudiantes que lograron éxito académico.

*Tasa de inflación*

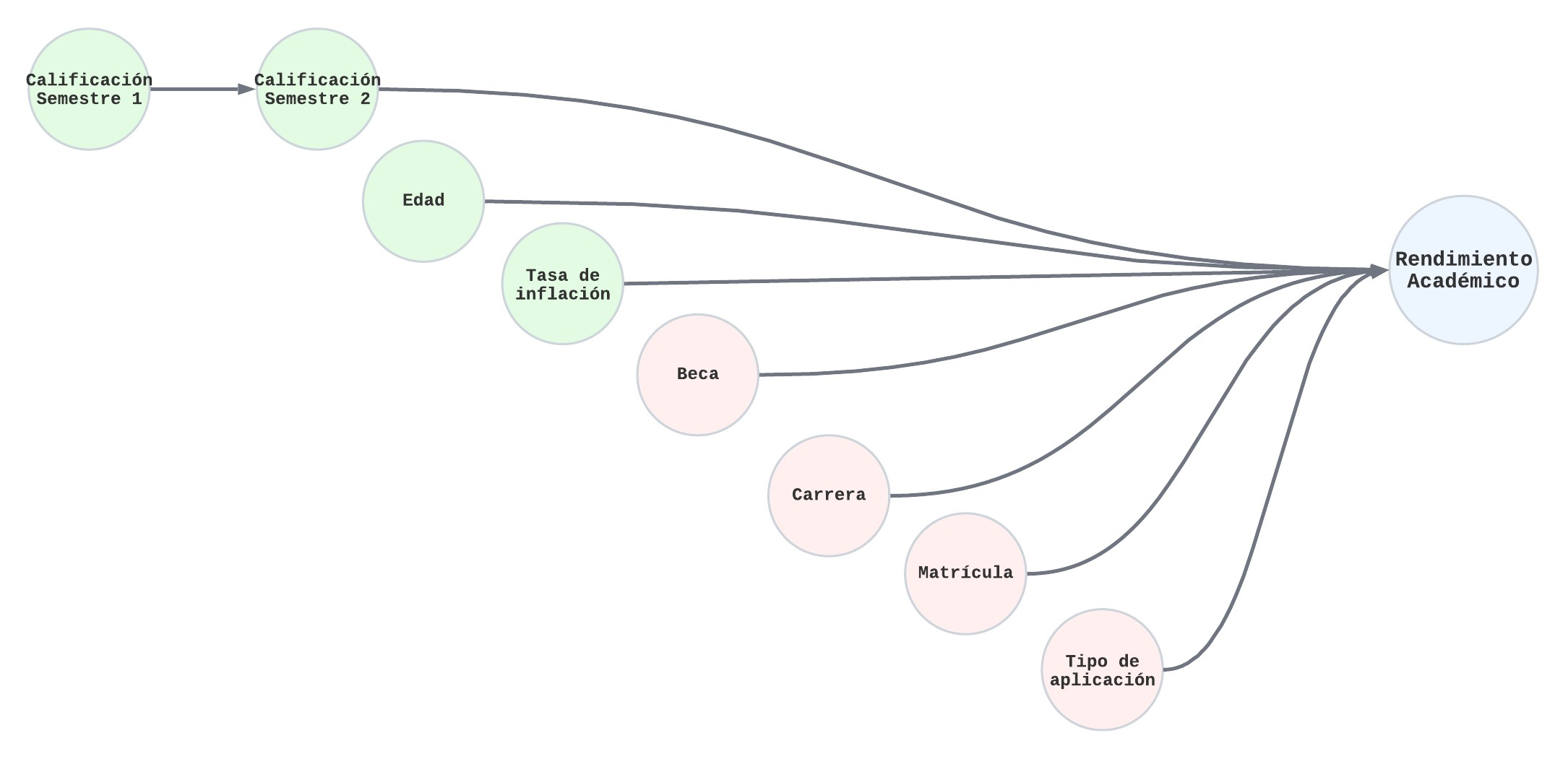


La tasa de inflación en ambos grupos presenta distribuciones sin diferencias significativas aparentes.

Una vez escogidas las variables se llevó a cabo una discretización de las variables continuas para poder ser utilizadas en la creación del modelo con redes bayesianas.

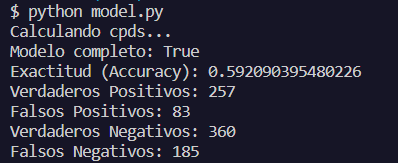
* La edad se redondeo al múltiplo de 5 más cercano
* Las notas se redondearon al múltiplo de 2 más cercano
* La tasa de inflación se redondeó al entero más cercano

Luego, se procedió a plantear la estructura de la red. Teniendo en cuenta que todas las variables tenían una alta correlación con el target, se conectaron directamente a dicha variable. No obstante, las notas de primer y segundo semestre se conectaron en cadena, considerando que, si se conoce la información de las notas de segundo semestre, las notas del primero dejarán de ser relevantes para la predicción porque se cuenta con información más actualizada y que en caso de que el estudiante aún no esté cursando segundo semestre, las notas del primero serán utilizadas. La estructura de la red se muestra a continuación:



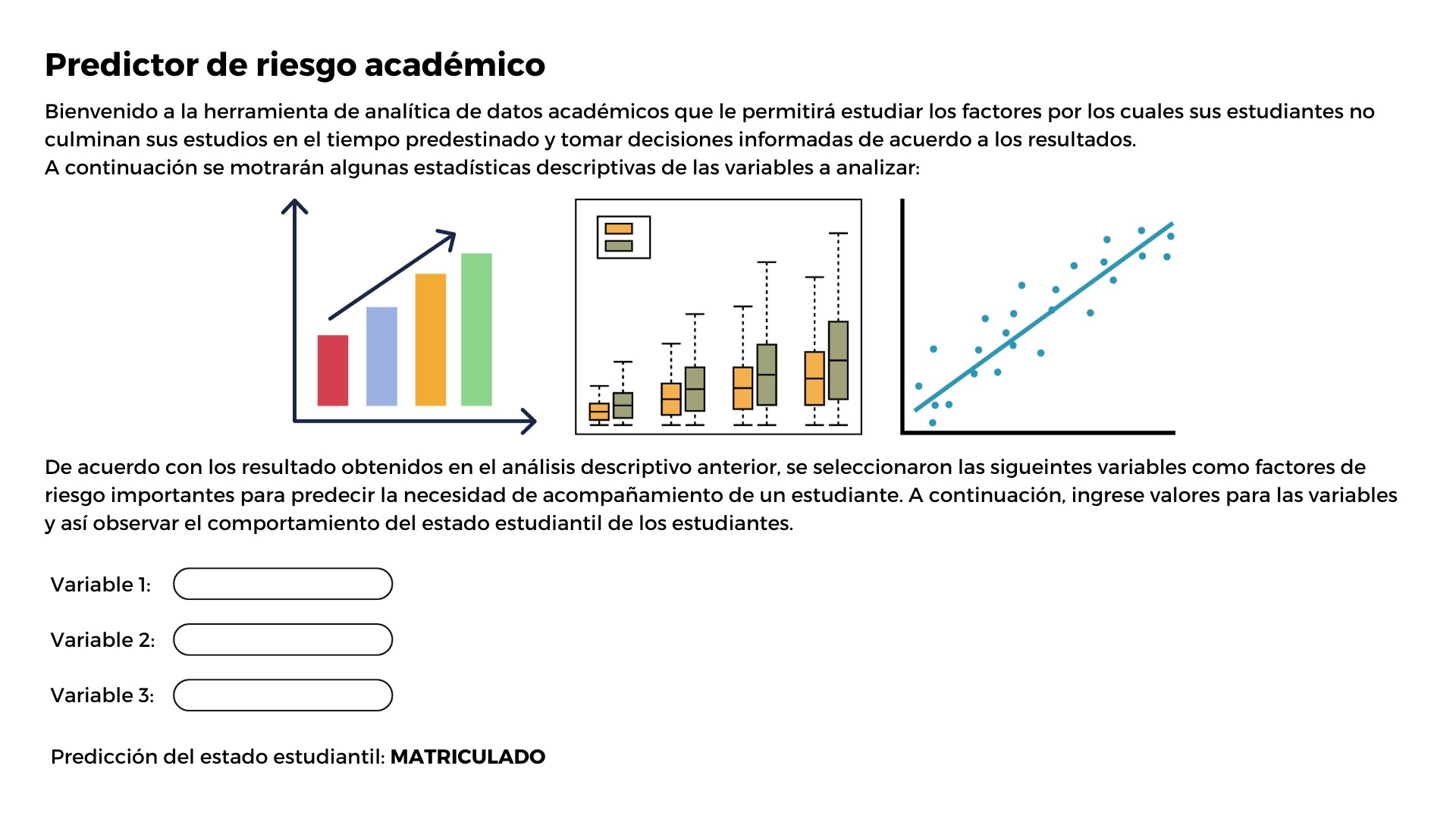
Una vez definida la red, se ajustaron las probabilidades de distribución condicionadas, se dividieron los datos en train (80%) y test (20%) considerando el desbalanceo de clases para que conservar la proporción en la cantidad de datos para cada clase en cada conjunto.

El modelo obtuvo los siguientes resultados:



Para obtener los resultados de Verdaderos Positivos, Verdaderos Negativos, Falsos Positivos y Falsos Negativos se consideró como positiva la clase de variación, y negativa la suma de los estudiantes pertenecientes a las clases seguir matriculado y desertar.

Para el dashboard primero se realizó un wireframe para tener una idea clara del tablero y los elementos que se pondrían en él, este diseño sencillo se muestra a continuación:



Posteriormente se decidió sólo mostrar las gráficas de correlación, ya que estas responden a la pregunta de investigación que hace referencia a cuáles variables pueden considerarse factores de riesgo para que un estudiante no tenga éxito académico en la universidad. Así mismo, al hacer la inferencia del estado estudiantil a partir de los datos suministrados por el usuario se busca responde a la pregunta de cómo predecir la necesidad de acompañamiento de un estudiante a partir de los factores de riesgo más relevantes.

El resultado final del tablero se muestra a continuación:

