Universidad Austral de Chile

Facultad de Ciencias de la Ingeniería

Escuela de Graduados

INFO 337

"ESTUDIO DE RETENCIÓN Y DESERCIÓN ESTUDIANTIL EN EDUCACIÓN SUPERIOR"

Profesores:
Eliana Scheihing García.
Yun Wang.

VALERIA NICOL SOTO CASTRO

VALDIVIA – CHILE

2024

INTRODUCCIÓN

1.1 Situación inicial

Las tasas de abandono de programas a nivel de educación superior, son una constante preocupación por parte de las Universidades. Persiste una búsqueda de factores de riesgo e intentos de probar posibles causas, sin lograr determinar variables predominantes, solo pruebas de buenos modelos de predicción. Es por esto la necesidad de identificar aquellos factores relevantes en la facultad de Ingeniería de la Universidad Austral, que a través de técnicas y herramientas de minería de datos ayuden a predecir, y la vez poder efectuar acciones efectivas frente a la problemática. Con el fin de evitar las altas tasas de abandono, y poder retener a los alumnos en sus programas de estudio.
Junto a esto, se suma un nuevo escenario que abre la posibilidad de cambios en indicadores de deserción, lo que fue la pandemia en los últimos dos años, por lo que es imprescindible estudiar situación pre y post pandemia.

Ante esta situación, se encuentran disponibles y serán tomados para análisis del trabajo, los datos académicos de los alumnos de la Facultad de Ingeniería, de la carrera de Ing. Civil Informática, de la Universidad Austral.

Motivo de esto se definen 3 preguntas de investigación para poder definir los objetivos del presente trabajo.

- ¿El fenómeno de deserción es distinto en tiempo de pandemia y pre pandemia?
- ¿Qué caracteriza la retención de estudiantes antes y durante la pandemia?
- ¿Qué estrategias son pertinentes para mejorar factores y apoyar la retención?

1.2 Objetivos

1.2.1 Objetivo General

Identificar las características de estudiantes retenidos y no retenidos, durante sus programas del primer año, en situación pre y durante pandemia, de la facultad de Ingeniería de la Universidad Austral de Chile, diseñando y evaluando modelos de predicción que permita apoyar en planes de intervención temprana en los estudiantes.

1.2.2 Objetivos Específicos

1. Definir enfoque metodológico con el que se desarrollará el trabajo, mediante la obtención de información y definición del set de datos.

2. Identificar fuentes de información, factores y variables de interés, mediante recolección de datos para posterior tratamiento de ellos.

3. Analizar diferentes técnicas y herramientas de minería de datos para predicciones.

4. Diseñar y ajustar modelos de predicción, para posterior evaluación de ellos.

5. Validar y concluir resultados.

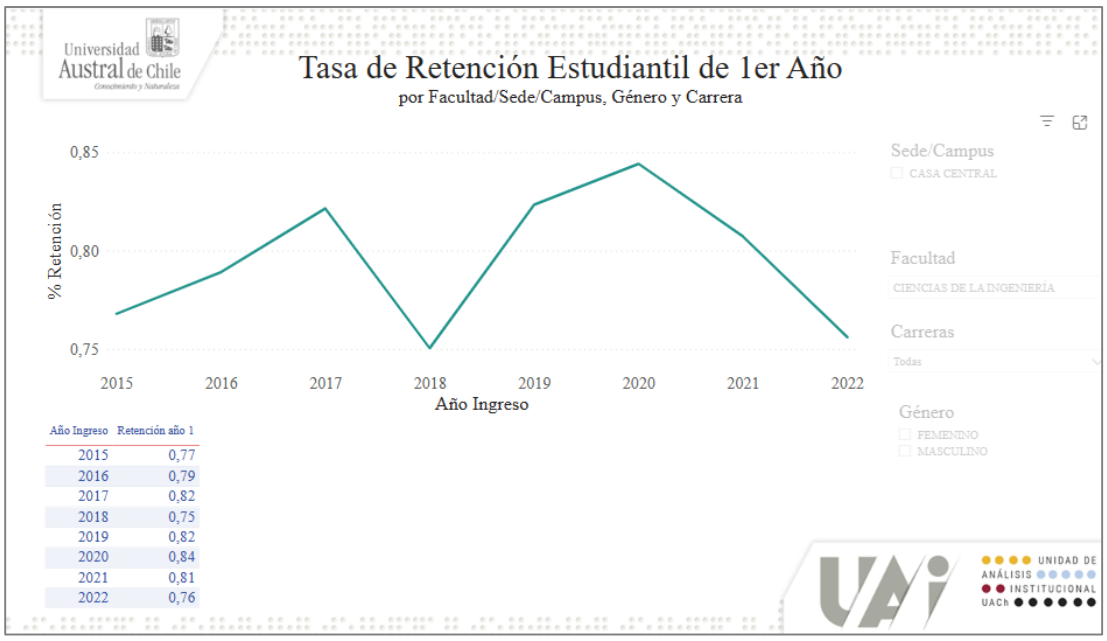
1.3 Trabajos relacionados y antecedentes bibliográficos

De acuerdo al último informe del SIES, del Ministerio de Educación a nivel nacional, la retención de primer año, es decir, los alumnos que se mantienen en la misma carrera, llegó a 75.5%, lo que significa una baja de 0,4% respecto el año anterior. La referencia se da respecto de los estudiantes de primer año de pregrado al 30 de abril de 2020 que continuaron matriculados al 30 de abril de 2021 en la misma institución.

Retención 1er Año	Año Cohorte 2016					
Subclasificación	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Universidades CRUCH	80,60%	81,50%	80,20%	80,70%	85,90%	84,20%
Universidades Privadas	76,90%	75,00%	76,60%	76,10%	82,40%	81,90%
Institutos Profesionales	68,30%	72,40%	71,90%	70,70%	70,10%	70,00%
Centros de Formación Técnica	65,20%	68,10%	69,40%	67,90%	66,10%	67,80%
Centros de Formación Técnica Estatales						66,70%
Total	**72,20%**	**74,20%**	**74,40%**	**73,70%**	**75,90%**	**75,50%**

Tabla 1: Retención Universitaria Chile 2016-2021.
Fuente: Elaboración propia, información plataforma SIES y CNED.

Ahondando en nuestro entorno, analizamos el contexto de la Universidad Austral de Chile, donde la Unidad de Análisis Institucional actualiza sus resultados públicos de análisis año a año, respecto a las tasas de retención. A continuación, presentamos el gráfico de interés, correspondiente a la Facultad de Ciencias de la Ingeniería de los últimos años:



Año Ingreso	Retención año 1
2015	0,77
2016	0,79
2017	0,82
2018	0,75
2019	0,82
2020	0,84
2021	0,81
2022	0,76

Gráfico 1: Tasa Retención Estudiantil UACh 2015-2022.
Fuente: Datos y análisis UAI-DTI.

Cabe destacar que la fecha de corte corresponde a 31 de diciembre de cada año, y 30 de abril de 2023 para año 2022. Respecto al primer año de retención. Por lo tanto, podemos inferir la notable baja en las tasas de retención en el año 2018, y el año 2022. Inflexiones que visualmente en gráfico, indican tendencia de cambios entre pre, durante y post pandemia.

Respecto a los factores relevantes que afectan a la deserción estudiantil, Chapman, D. W., & Pascarella, E. T. (1983), nos indican una serie de factores a considerar para explicar la deserción estudiantil universitaria. Como: género, raza, tipo de colegio secundario, desempeño académico, ingreso académico, entre otros. Por lo que es útil asociarlos y clasificarlos de la siguiente manera:

Demográficos	Socioculturales	Socio económicos	Factor académico
•Genero •Edad •Residencia	•Enseñanza media (lugar estudios) •Aspectos motivacionales	•Beneficios financiamiento académico •Nivel socio económico familiar	•Ranking •PSU •NEM •Carrera pregrado que ingresó primer año •**Rendimiento estudiante ***

Tabla 2: Factores asociados a la deserción y retención estudiantil.
Fuente: Elaboración propia.

1 DESARROLLO DEL TRABAJO

1.1 Metodología

A continuación, en Figura 1 se presentan las actividades relacionadas para llevar a cabo el cumplimiento de los objetivos mencionados anteriormente para el desarrollo del trabajo. En primera instancia se basa en la investigación de información y obtención de datos cualitativos y cuantitativos. Posterior a esto, se desarrolla análisis exploratorio de los datos, para obtener como resultado las variables a definir, métricas y análisis de correlación respectivos. Siguiente de esto, se recurre al desarrollo y utilización de métodos y herramientas estadísticas, para obtener apoyo en resultados efectivos, desarrollo y ajuste de modelos a utilizar. Finalmente, se espera comparar modelos, validar resultados obtenidos, y concluir el efecto obtenido, en la carrera de Ingeniería Civil Informática, Facultad de Ingeniería UACh.

Importante mencionar, que el trabajo se desarrolla a través de Jupyter Notebook, con respectivos lenguajes R y Python, más librerías necesarias según modelos.

A continuación, se presenta bosquejo resumen del trabajo metodológicamente:

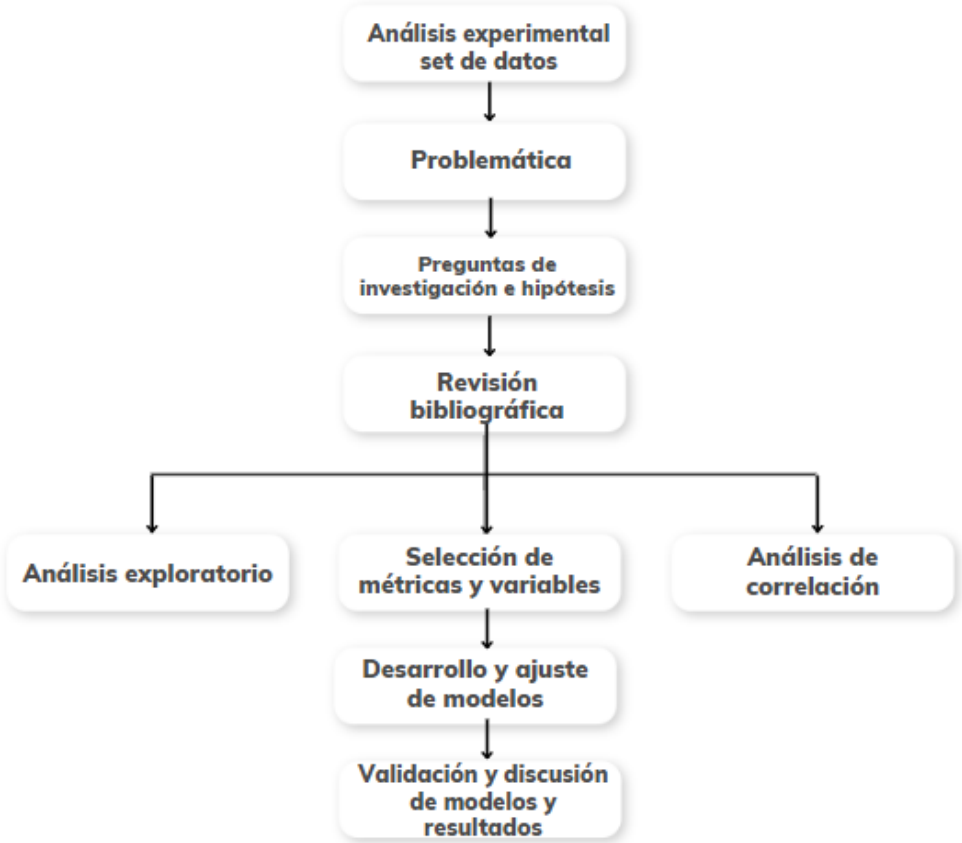


Figura 1: Metodología.

Para las técnicas y herramientas estadísticas desarrolladas, se empleó una serie de pasos, donde se definió la siguiente metodología para llevar a cabo estas tareas sobre los datos obtenidos:

- Preparación de datos.
- Instalación de paquetes.
- Análisis Exploratorio.
- Desarrollo de modelos.
- Ajuste de modelos a datos.
- Evaluación de modelos.
- Interpretación de resultados.
- Predicciones.

1.1 Preparación de datos

Los datos analizados corresponden a estudiantes de la carrera de Ingeniería Civil Informática de la Universidad Austral de Chile. En la recolección de datos, se obtienen parámetros según semestres, respecto a notas, asignaturas cursadas y no cursadas, créditos respectivos y años correspondiente de entrada y estado actual, es decir, de aspecto de rendición académica.

Posteriormente se identificaron aquellos estudiantes que permanecieron en su tercer semestre y/o tuvieron matrícula vigente, siendo categorizados con valor 1; y con valor 0 aquellos que no permanecieron. Finalmente se definen dos grupos de datos para clasificar el tratamiento de ellos, siendo primer grupo aquellos datos correspondientes a los años 2017 y 2018, aquel grupo definido como “pre pandemia”; y el segundo grupo, correspondiente a los años 2020 y 2021, aquel grupo definido como “durante pandemia”. Categorizándose como valor 0 el primer grupo, y valor 1 el segundo grupo.

Estos datos son ordenados en varios features según semestre 1 y semestre 2, siendo la etiqueta de decisión de permanencia, si se matriculó en el primer semestre del segundo año o no.

Basado en esto, se definieron las siguientes variables para el posterior desarrollo de modelos:

- Variable Tratamiento = year
- Variable Respuesta = Permanencia_term_3
- X1 = total_credits_1
- X2 = total_credits_2
- X3 = total_courses_1
- X4 = total_courses_2
- X5 = course_approved_1
- X6 = course_approved_2
- X7 = course_failed_1
- X8 = course_failed_2
- X9 = t_gpa_1

- X10 = t_gpa_2
- X11 = c_gpa_1
- X12 = c_gpa_2

1.2 Análisis exploratorio

Para comenzar, se realizó lectura de los datos ya procesados anteriormente, a través de Jupyter Notebook.

Se realiza un análisis exploratorio de datos, donde se identifican 286 datos, es decir, 286 estudiantes a analizar, y 18 columnas, las cuales corresponden a variables de predicción, tratamiento y respuesta.

En primer lugar, se desarrolló el cálculo de correlaciones, a través de una matriz de correlación, para poder interpretar la relación entre las variables con las que contamos en los datos. Indicador relevante para posterior selección y/o evaluación de ellas.

En segundo lugar, en apoyo a definición de variables, es desarrollado el modelo de regresión binaria probit, definido anteriormente, donde se toman un total de 10 variables, para ver la significancia y comportamiento de ellas.

En conjunto, con el fin de analizar más detalladamente estas variables, se añadieron gráficos de dispersión de apoyo para evaluar el comportamiento de cada una, correspondiente a las probabilidades predichas, y el modelo probit. Por último, se incorpora un análisis PCA para complementar el comportamiento de la data y variables.

1.3 Desarrollo de modelos

Si bien se espera destaquen y analizar variables significativas en el punto anterior, se define un ajuste al modelo de regresión probit, desarrollando el modelo BART, descrito anteriormente, con el fin de poder encontrar mejoras en los resultados.

Importante mencionar, que la descripción del modelo BART se basa en Hugh A. Chipman, Edward I. George And Robert E. Mcculloch (2010).

1.3.1 Modelo BART

Consideramos el problema fundamental de hacer inferencias sobre una función desconocida f que predice una salida Y usando un vector dimensional p de entradas x cuando,

(1) $$Y = f(x) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Para realizar esto, consideramos modelar o al menos aproximar f (x) = E(Y |x), la media de Y dado x, mediante una suma de m árboles de regresión f (x) ≈ h(x) ≡ m j=1 gj (x) donde cada gj denota un árbol de regresión. Por lo tanto, aproximamos (1) por un modo de suma de árboles:

(2) $$f(x) \approx h(x) \equiv \sum_{j=1}^{m} g_j(x)$$

Donde finalmente la ecuación se actualiza a:

(3) $$Y = h(x) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Respecto a la elaboración de la forma del modelo de suma de árboles (2), comenzamos por establecer la notación para un modelo de árbol único. Sea T un árbol binario que consta de un conjunto de reglas de decisión de nodos interiores y un conjunto de nodos terminales, y sea M = {μ1,μ2,...,μb} un conjunto de valores de parámetros asociados con cada uno de los b terminales nodos de T. Las reglas de decisión son divisiones binarias del espacio predictor de la forma {x ∈ A} vs {x /∈ A} donde A es un subconjunto del rango de x. Por lo general, se basan en los componentes individuales de x =(x1,...,xp) y tienen la forma {xi ≤ c} frente a {xi > c} para xi continuo. Cada valor de x está asociado con un solo nodo terminal de T por la secuencia de reglas de decisión de arriba a abajo, y luego se le asigna el valor de μi asociado con este nodo terminal. Para T y M dados, usamos g(x; T,M) para denotar la función que asigna un μi ∈ M a x.

(4) $$Y = \sum^{m} g(x; T_j, M_j) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Finalmente, se completa la especificación del modelo BART imponiendo una prioridad sobre todos los parámetros del modelo de suma de árboles quedando la ecuación final (4).

Bajo tales prioridades, los componentes del árbol (Tj ,Mj) son independientes entre sí y de σ , y los parámetros del nodo terminal de cada árbol son independientes. Las restricciones de independencia anteriores simplifican el problema de especificación previa a la especificación de formas solo para p(Tj),p(μij |Tj) y p(σ), una especificación que simplificamos aún más usando formas idénticas para todas las p(Tj) y para todo p(μij |Tj).

(5) $$p(M_j|T_j) = \prod_i p(\mu_{ij}|T_j).$$

Teniendo en cuenta el modelo final (4) descrito anteriormente donde:

- T es un árbol binario;
- M = (μ1, μ2, .., μb) es el vector de medias en los nodos terminales b del árbol;
- g(z, x; T, M): valor obtenido siguiendo la observación (z,x) árbol abajo y devolviendo la media para el nodo terminal en el que aterriza;
- E ~ N (0, σ^2);

Se procede a desarrollar el modelo en lenguaje R, a través de la librería “BayesTree”, paquete instalado previo a ejecutar modelo. Y en conjunto, se generan gráficos que muestran cómo cambian las predicciones del modelo

BART a medida que varía una variable en particular, manteniendo constantes todas las demás.

1.3.2 Especialización modelo BART C

Para este modelo, se toma misma definición anterior, más nuestro efecto de causalidad. Se toman en cuenta mismas variables definidas en el modelo BART, realizando el ajuste enfocado en BART C. Donde toma por defecto, algoritmo estudiado en artículo de Hill (2011).

A través de la librería "bartCause", paquete instalado en R, se procede a desarrollar este ajuste en el modelo, donde sumamos un "z" el cual corresponde a un tratamiento binario, que considera el efecto causal.

En esta instancia, se espera analizar principalmente los resultados estadísticos de esta adaptación.

2 RESULTADOS

En esta sección, se describen los resultados obtenidos, según la metodología definida y desarrollada anteriormente.

2.1 Análisis exploratorio

A continuación, en figura 2, con el fin de ver las distribuciones de las variables, principalmente aquellas numéricas, se muestra el resumen estadístico obtenido de los datos, para tener en cuenta métricas importantes de cada una:

```
      id                                  student_id       year
Min.   :  1.00   00b766201f72e60666f22c02317a5d:  1   Min.   :0.0000
1st Qu.: 72.25   00fe221fac0f4d1bf2683dcd7e3d66:  1   1st Qu.:0.0000
Median :143.50   0382ead66419e1409a953d0500f790:  1   Median :1.0000
Mean   :143.50   0419b0933c02c07b7f6be7a58d3b62:  1   Mean   :0.5105
3rd Qu.:214.75   047594087b6b6ce60f7cb0bc405a23:  1   3rd Qu.:1.0000
Max.   :286.00   04d40c955dd809a1dd13b0f6e5c065:  1   Max.   :1.0000
                 (Other)                        :280
total_credits_1 total_credits_2 total_courses_1 total_courses_2
Min.   : 9.0    Min.   : 0.00   Min.   :2.000   Min.   :0.000
1st Qu.:28.0    1st Qu.:24.00   1st Qu.:5.000   1st Qu.:5.000
Median :28.0    Median :31.00   Median :5.000   Median :6.000
Mean   :27.4    Mean   :26.21   Mean   :4.825   Mean   :5.434
3rd Qu.:29.0    3rd Qu.:31.00   3rd Qu.:5.000   3rd Qu.:7.000
Max.   :30.0    Max.   :36.00   Max.   :6.000   Max.   :8.000

course_approved_1 course_approved_2 course_failed_1 course_failed_2
Min.   :0.000     Min.   :0.000     Min.   :0.00    Min.   :0.000
1st Qu.:2.000     1st Qu.:2.000     1st Qu.:0.00    1st Qu.:0.000
Median :3.000     Median :4.000     Median :1.00    Median :1.000
Mean   :3.395     Mean   :3.846     Mean   :1.43    Mean   :1.587
3rd Qu.:5.000     3rd Qu.:6.000     3rd Qu.:3.00    3rd Qu.:3.000
Max.   :5.000     Max.   :7.000     Max.   :5.00    Max.   :7.000

   t_gpa_1         t_gpa_2         c_gpa_1         c_gpa_2        start_year
Min.   :1.020   Min.   :0.000   Min.   :1.020   Min.   :0.000   Min.   :2017
1st Qu.:3.692   1st Qu.:3.000   1st Qu.:3.712   1st Qu.:3.865   1st Qu.:2017
Median :4.410   Median :4.255   Median :4.460   Median :4.565   Median :2020
Mean   :4.370   Mean   :3.744   Mean   :4.387   Mean   :4.101   Mean   :2019
3rd Qu.:5.025   3rd Qu.:4.810   3rd Qu.:5.025   3rd Qu.:4.985   3rd Qu.:2021
Max.   :6.610   Max.   :6.540   Max.   :6.610   Max.   :6.570   Max.   :2021

   Cluster      Permanencia_term_3
Min.   :1.000   Min.   :0.0000
1st Qu.:1.000   1st Qu.:1.0000
Median :2.000   Median :1.0000
Mean   :1.867   Mean   :0.7552
3rd Qu.:3.000   3rd Qu.:1.0000
Max.   :3.000   Max.   :1.0000
```

Figura 2: Resumen estadístico.

Según resultados interpretados de matriz de correlación, en figura 4, podemos observar que ciertas variables tienen mayor relación entre ellas y el efecto que causarían en la “respuesta”. Es por ello, que basado en este análisis, donde según intensidad de color y tamaños del círculo, es la proporción al factor de correlación, se escogieron y definieron variables predominantes para el primer modelo desarrollado. Cabe destacar que la matriz de correlación principalmente nos indica que no existe correlación entre la variable tratamiento y variable respuesta.

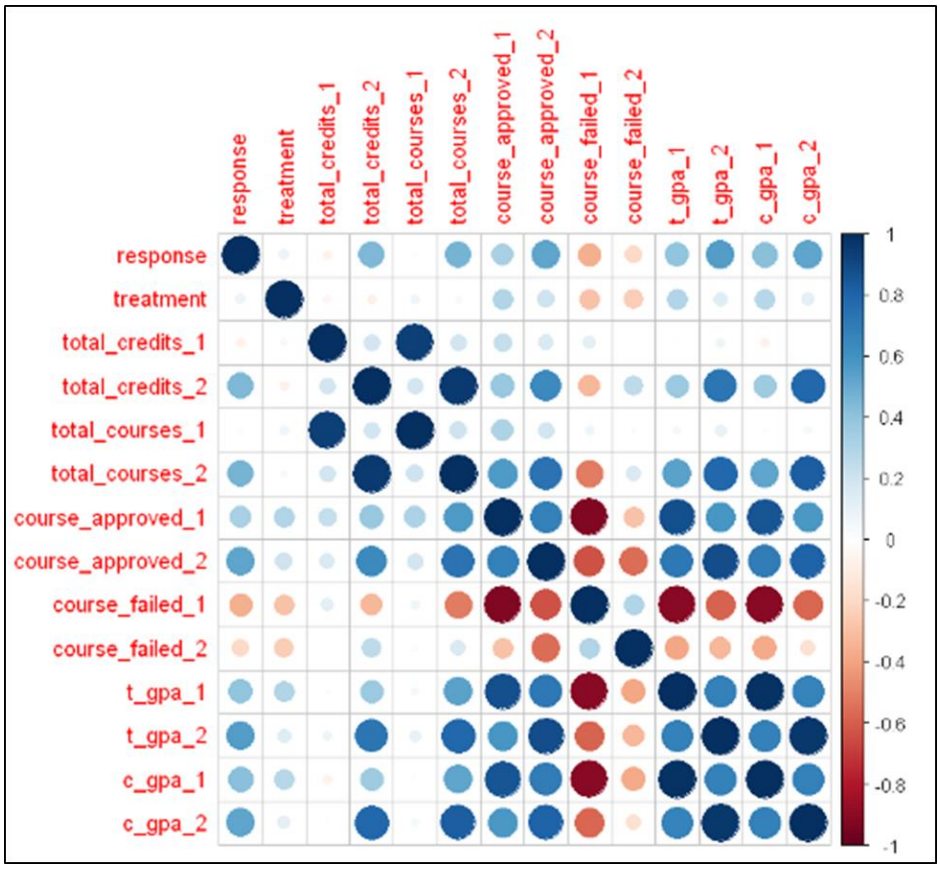


Figura 3: Matriz de correlación.

Basado en esto, y en elección propia de variables predominantes a evaluar, fue óptimo seleccionar las siguientes variables predictoras:

- Total créditos 1° Semestre
- Total créditos 2° Semestre
- Total cursos 2° Semestre
- Total cursos aprobados 1° Semestre
- Total cursos aprobados 2° Semestre
- Total GPA 1° Semestre
- Total GPA 2° Semestre

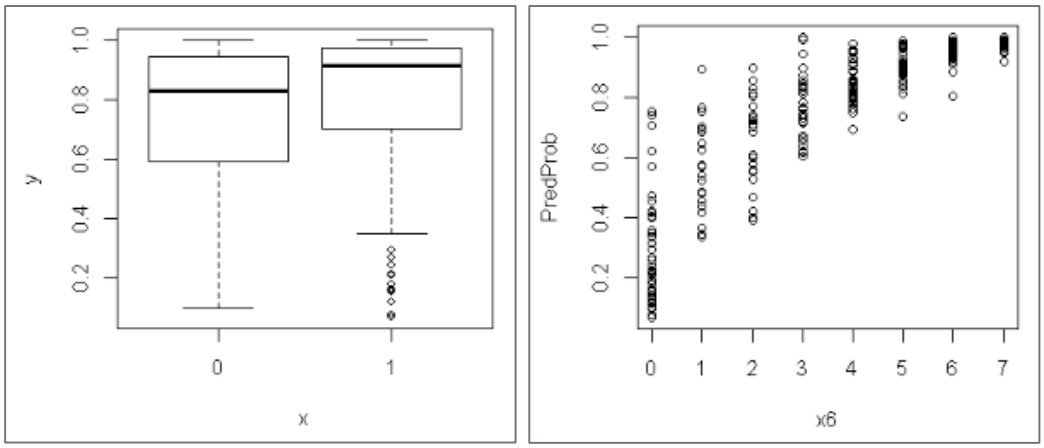
De esta forma obtuvimos el siguiente resultado:

```
Call:
glm(formula = y ~ z + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 +
    x9 + x10, family = binomial(link = "probit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.27860   0.03716   0.27610   0.59503   1.93992

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.397544   1.486340  -0.267  0.78911
z1          -0.009755   0.215768  -0.045  0.96394
x1          -0.218624   0.093158  -2.347  0.01894 *
x2           0.136523   0.055595   2.456  0.01406 *
x3           0.923588   0.554568   1.665  0.09583 .
x4          -0.596606   0.277164  -2.153  0.03136 *
x5          -0.016751   0.207582  -0.081  0.93568
x6           0.309298   0.097702   3.166  0.00155 **
x7                 NA         NA      NA       NA
x8                 NA         NA      NA       NA
x9           0.295457   0.278403   1.061  0.28857
x10          0.039392   0.145451   0.271  0.78652
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318.32  on 285  degrees of freedom
Residual deviance: 212.36  on 276  degrees of freedom
AIC: 232.36

Number of Fisher Scoring iterations: 6
```

Figura 4: Modelo regresión binaria probit.

Podemos observar de acuerdo a estos resultados que aquellas variables significativas serían las variables X1, X2, X4 y X6.

Luego de esto, como se mencionó anteriormente, se realizaron los gráficos de dispersión para cada una de estas variables significativas, de acuerdo a modelo recién desarrollado, los cuales se presentan a continuación:



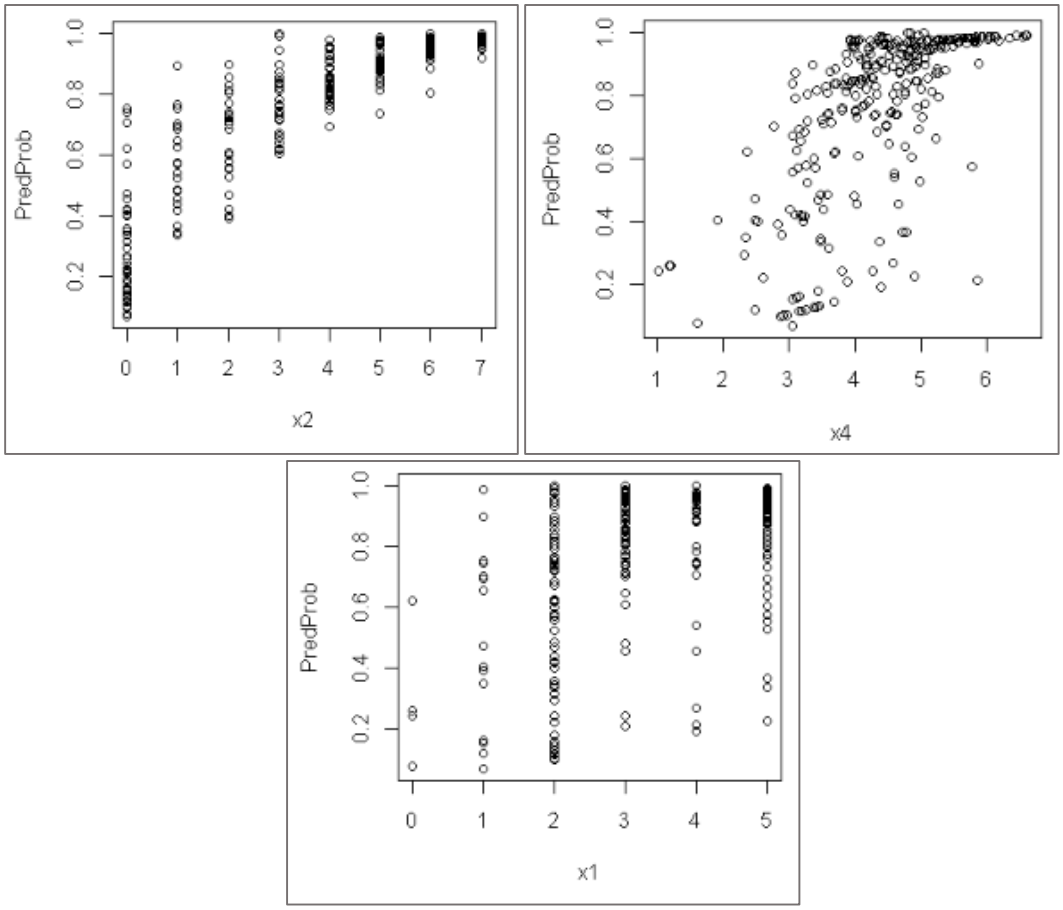


Figura 5: Gráficos de dispersión modelo probit y variables.

Analizando cada uno de estos gráficos, se observó que las variables con mayor significancia serían principalmente la X2 y X6, ya que presentaron una clara tendencia, consistencia y curva en su dispersión, lo que me indica que, si los valores de estas variables independientes aumentan, las probabilidades predichas también tienden a aumentar.

De la misma forma, complementando el análisis de datos y variables, en figura 7, podemos ver el resultado del análisis PCA que se desarrolló. Visualmente se infiere según las cargas representadas entre las variables y los componentes principales, aquellas que tienen mayor influencia, como lo son las variables X1, X2, X4, X5, y X3. Estando relacionadas las variables X2 y X5 positivamente correlacionadas según dirección; la X4 y X1 con correlación negativa según su dirección, y para el caso de X3 no tendría agrupación con otra y tendría correlación positiva con su componente.

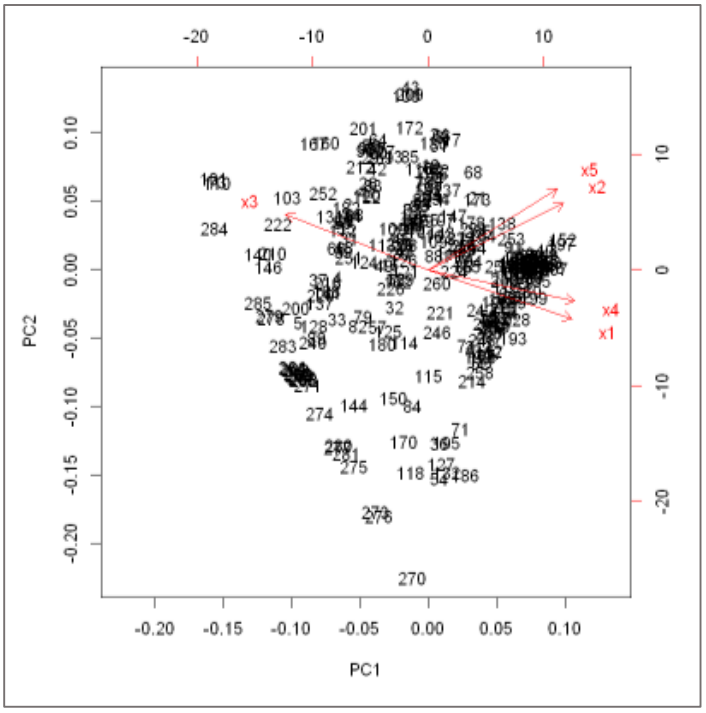


Figura 6: PCA data.

2.2 Desarrollo de modelos

2.2.1 BART

Luego de ejecutar el modelo BART definido en punto anterior, donde se tomó el prior por defecto de k=2, y un número de árboles total de 200; se obtuvieron los siguientes resultados e iteraciones:

```
number of trees: 200
Prior:
        k: 2.000000
        binary offset is: 0.000000
        power and base for tree prior: 2.000000 0.950000
        use quantiles for rule cut points: 0
data:
        number of training observations: 239
        number of test observations: 47
        number of explanatory variables: 5


Cutoff rules c in x<=c vs x>c
Number of cutoffs: (var: number of possible c):
(1: 100) (2: 100) (3: 100) (4: 100) (5: 100)
```

```
Running mcmc loop:
iteration: 100 (of 1100)
iteration: 200 (of 1100)
iteration: 300 (of 1100)
iteration: 400 (of 1100)
iteration: 500 (of 1100)
iteration: 600 (of 1100)
iteration: 700 (of 1100)
iteration: 800 (of 1100)
iteration: 900 (of 1100)
iteration: 1000 (of 1100)
iteration: 1100 (of 1100)
time for loop: 10

Tree sizes, last iteration:
2 2 3 3 2 2 2 2 2 2 3 2 2 2 2 2 2 3 3 2
3 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 3 2
2 2 3 3 3 3 1 2 2 2 2 2 2 2 2 4 2 4 3 1
2 2 2 1 2 3 2 2 2 2 2 2 1 3 2 2 3 2 3 2
3 2 2 3 3 2 3 2 2 2 2 3 2 2 3 2 2 2 2 2
2 4 2 2 3 2 2 2 3 3 2 3 2 3 2 4 3 2 1 2
2 2 2 3 4 2 2 2 2 2 2 2 4 2 2 2 2 3 2 2
2 3 2 2 2 2 3 2 3 2 2 2 2 3 4 5 2 2 4 2
2 1 2 4 1 3 3 3 3 3 2 2 3 2 2 3 2 5 1 2
2 2 2 2 2 2 2 2 2 3 2 2 2 5 3 3 2 2 2 2
Variable Usage, last iteration (var:count):
(1: 52) (2: 61) (3: 48) (4: 51) (5: 53)

DONE BART 11-2-2014
```

Figura 7: Resultados e iteraciones modelo BART.

Seguido de esto, se procede a generar gráficos para ver la variación del modelo y variables. En este caso fue para las variables X1, X2, X3, X4 y X5, que fueron las definidas a utilizar en este modelo, correspondiente a:

X1: total_credits_2
X2: total_courses_1
X3: total_courses_2
X4: course_approved_2
X5: course_failed_2

De esta forma, se obtuvieron los gráficos que se muestran en figura 9, donde vemos cómo cambian las predicciones del modelo BART a medida que varía una variable en particular, manteniendo constantes todas las demás.

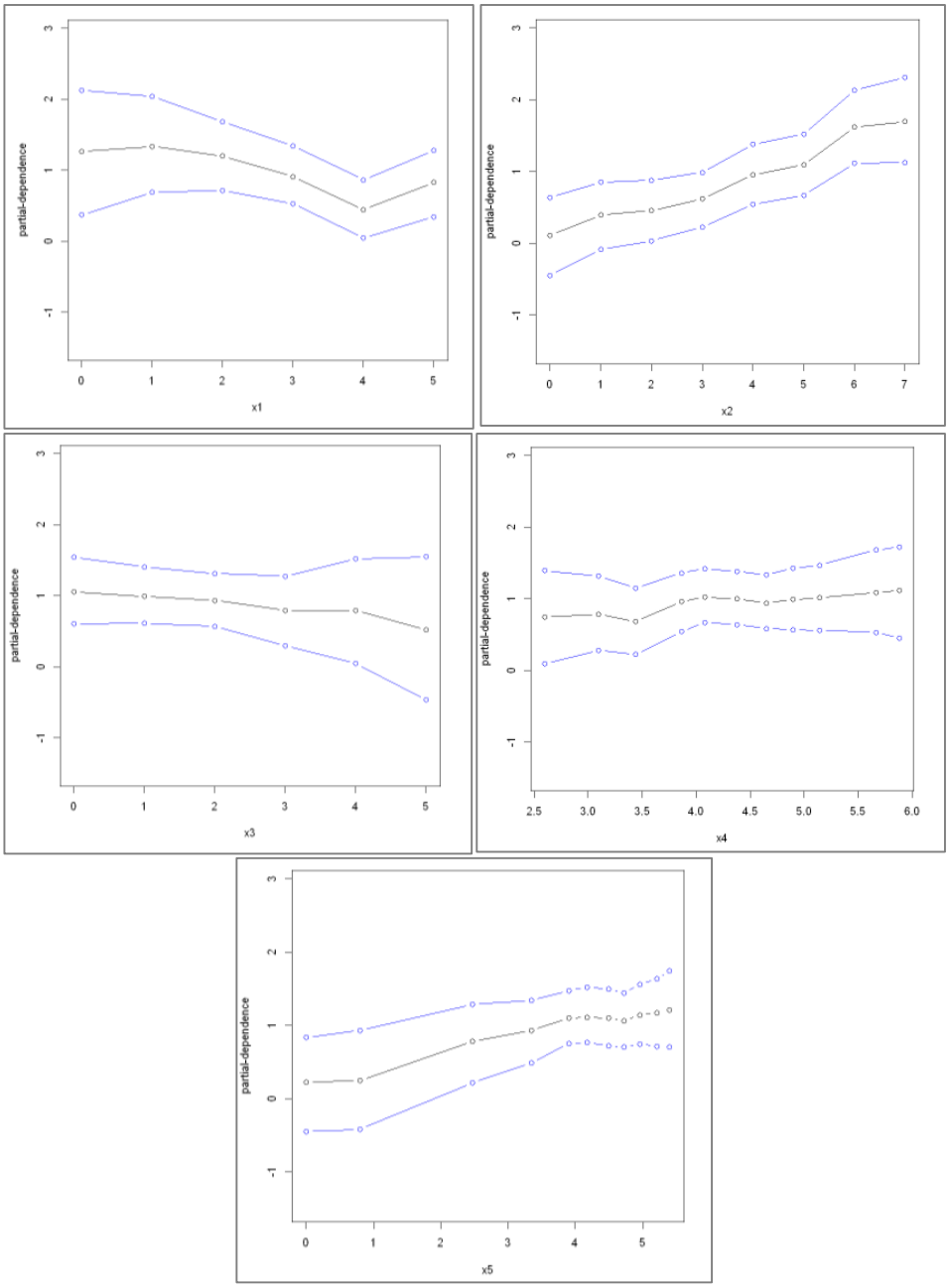


Figura 8: Gráficos modelo BART y variables.

En este caso el enfoque de nuestro análisis a través de los gráficos generados por cada variable categórica, es la representación de cómo cambian las predicciones del modelo, en rangos específicos de valores de la variable independiente, por eso podemos observar cómo se muestran diferentes intervalos, y como puede cambiar la tendencia entre ellos, para una sola variable.

Basado en esto, se observa que las variables con tendencia significativa son las variables X2 y X5.

2.2.2 Especialización modelo BART C

Posteriormente, se desarrolló la especialización BART C, descrita anteriormente, y se obtuvo el siguiente resultado:

```
Treatment effect (population average):
    estimate      sd ci.lower ci.upper
ate -0.07626 0.09698  -0.2663   0.1138
Estimates fit from 286 total observations
95% credible interval calculated by: normal approximation
  population TE approximated by: posterior predictive distribution
Result based on 100 posterior samples times 2 chains
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.328671 -0.143357 -0.069930 -0.074965 -0.003497  0.129371
```

Figura 9: Resultados BART C.

Estos resultados, me indican la baja relación que existe en el efecto del tratamiento, según el modelo. Ya que, según el intervalo de confianza obtenido, el cual incluye el uno, nos indica que no está siendo un modelo predictivo en estos datos, específicamente.

2.3 Comparación de modelos

Con el fin de evaluar la precisión y rendimiento de las predicciones de los 3 modelos desarrollados, se traza la sensibilidad frente a la especificidad de una prueba, en curvas ROC. Para poder comparar, se trazan las curvas del modelo Regresión binaria probit, BART y BART C, junto a la línea de no discriminación, es decir, aquella sin entrenar.

El resultado fue el siguiente gráfico de curvas:

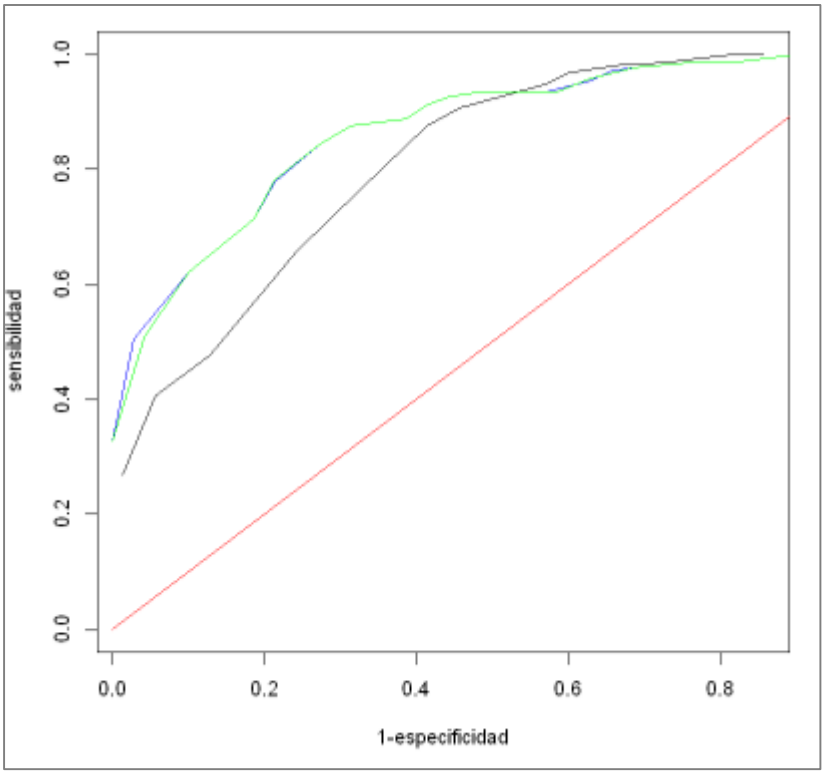


Figura 10: Curvas ROC.

De acuerdo a la interpretación de sensibilidad de las curvas ROC, estas me indican que aquellas curvas que se presenten en el área bajo la línea de no discriminación serían resultados negativos, o peor de una curva, mientras que aquella que se encuentre en el área sobre la línea de no discriminación, serían de buen resultado y mejores.

Por lo tanto, según las curvas obtenidas, podemos ver que, para los datos analizados, los 3 modelos serían significativos, es decir, nos entregan una buena predicción. Sin embargo, los modelos BART y BART C (color verde y azul), serían muy similares en sus resultados y mejores que el modelo de regresión binaria probit (color negro).

Conjunto a esto, en el análisis final de variables significativas, teniendo como las más predominantes en la causa de la retención y deserción del estudiante, fueron:

- Total cursos 1° Semestre.
- Cursos fallido 2° Semestre.
- Cursos aprobados 2° Semestre.
- Total créditos 2° Semestre.

3 DISCUSIÓN Y CONCLUSIONES

En primera instancia, es importante destacar, que el origen de elegir trabajar con árboles de decisión, fue basado en la necesidad de trabajar con efectos de causalidad posteriormente.

En respuesta a las preguntas de investigación definidas; en primer lugar, si bien no se definen características claras de los estudiantes no retenidos, pudimos ver que aquellos factores que influyen en estos datos y análisis, con mayor significante y de acuerdo al efecto del resto de variables, son principalmente los cursos aprobados.

Finalmente, según resultados obtenidos, no existe efecto directo de la pandemia, en la deserción y retención de estudiantes, de la carrera de Ing. Civil Informática UACh. Aparentemente, quien podría entregar un efecto positivo en ello, es la flexibilidad entregada en periodo de pandemia.

Sin embargo, es notorio los puntos de inflexión encontrados en periodos de pre pandemia, durante pandemia, y post pandemia, que corresponden al año actual; por lo que sería recomendable e interesante poder incluir estos nuevos datos en el análisis, para ver qué ocurre con los resultados post pandemia.

Adicionalmente, la presencia de una tendencia que muestra inicialmente una disminución seguida de un posterior crecimiento en las tasas de retención, sugiere la existencia de factores influyentes entre los estudiantes. Por ende, resulta de suma importancia la incorporación de nuevas variables que sean capaces de proporcionar un análisis más completo sobre el efecto causal en su retención. Estas variables no deben limitarse exclusivamente a los relacionados con el rendimiento académico, sino que deben abarcar una gama más amplia de características.

Por lo tanto, se espera aumentar campo de datos y variables, para poder probar los modelos BART y BART C principalmente, debido a su buen rendimiento tomando en cuenta la causalidad, y así colaborar de forma efectiva con personas involucradas en el plan estratégico de la facultad y universidad.

Por último, importante mencionar, que ya se encuentran en proceso recibir en bruto datos actualizados de alumnos de la Facultad de Ingeniería de la Universidad Austral para análisis en año 2024, donde es posible encontrar nuevos resultados satisfactorios luego del tratamiento de ellos. Esto debido a la disposición de nuevas variables del estudiante, como puntajes psu, número de preferencia de carrera en postulación, ciudad de origen, notas enseñanza media, entre otras, que se infiere nos abrirán nuevos resultados, por lo que los pasos a seguir es comenzar a trabajar en ellos y posterior desarrollo de modelos y ajustes, definiendo si es necesario nuevas preguntas de investigación y adecuación de objetivos.

4 REFERENCIAS

[1] Jennifer L. Hill, (2010), Bayesian Nonparametric Modeling for Causal Inference.

[2] Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). Introduction to the Theory of Statistics. McGraw-Hill Education.

[3] Hugh A. Chipman, Edward I. George And Robert E. Mcculloch (2010). Bart: Bayesian Additive Regression Trees.

[4] Himmel, E. (2002). Modelo de análisis de la deserción estudiantil en la educación superior.

[5] OEA (2003). Documento Base del Proyecto. Estrategias y Materiales Pedagógicos para la Retención Escolar
[6] Cortés Gallardo, N., Aguilar Vivar, A., González Mimica, M., & Muñoz Jeréz, Z. (2016). Perfil Del Estudiantes Que Deserta De La Universidad Austral De Chile Entre Los Años 2010-2015.
[7] Agrusti F., Bonavolontà G., Mezzini M.(2019),Predicción de la deserción universitaria a través de técnicas de minería de datos: una revisión sistemática.
[8] Von Hippel, Paul T. y Alvaro Hofflinger. (2020). La revolución de los datos llega a la educación superior: Identificando estudiantes en riesgo de deserción en Chile.
[9] Salazar-Fernández, JP; Sepúlveda, M.; Muñoz-Gama, J.; Nussbaum, M. Análisis curricular para caracterizar trayectorias educativas en cursos con alta tasa de reprobación que conducen a la deserción tardía.
[10] Informe retención pregrado SIES 2022.