UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MODELO DE PROPENSION DE MATRÍCULA DE ALUMNOS CONVOCADOS EN UNA UNIVERSIDAD ADSCRITA AL SISTEMA ÚNICO DE ADMISIÓN (SUA)

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

RUPERTO ALEJANDRO ORTIZ VARGAS

PROFESOR GUÍA

CAROLINA SEGOVIA RIQUELME

MIEMBROS DE LA COMISIÓN

RICHARD WEBER HAAS

RODRIGO MORALES LAVANDEROS

SANTIAGO DE CHILE
2019

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR: RUPERTO ORTIZ VARGAS
FECHA: 15/07/2019
PROFESOR GUÍA: CAROLINA SEGOVIA

MODELO DE PROPENSION DE MATRÍCULA DE ALUMNOS CONVOCADOS EN UNA UNIVERSIDAD ADSCRITA AL SISTEMA ÚNICO DE ADMISIÓN (SUA)

Cada año las personas interesadas en acceder a la educación superior chilena, deben inscribirse dentro del sistema único de admisión universitaria (SUA). En este participan 41 universidades, entre las que se encuentran las universidades adscritas al CRUCH y algunas universidades privadas. La Universidad en estudio es la más grande a nivel nacional en términos de alumnos y ha sido la universidad que históricamente ha acaparado el mayor número de postulaciones, sin embargo, ha tenido dificultades a la hora de completar sus cupos, pues solamente cerca del 60% de los alumnos que quedan seleccionados en esta casa de estudios finaliza su matrícula en el primer periodo establecido por el DEMRE.

El objetivo principal de este trabajo de título es desarrollar un modelo que permita clasificar a los alumnos convocados con menor tendencia a concretar su matrícula, de forma tal que se puedan tomar acciones que permitan aumentar los chances de que el convocado se matricule durante el 1° proceso de matrículas.

Se comparan las respuestas de los modelos propuesto en una campaña de contacto simulada. Los escenarios utilizan la información del 2019 para la evaluación. Se utiliza como métricas de evaluación los costos esperados y beneficios extra entregados por cada modelo en escenarios de alta, media y bajas tasas de respuesta positiva a matricularse por parte de los convocados. Los costos y el análisis de sensibilidad de los resultados están basado en componentes del negocio. Dentro de todos los modelos propuestos, la regresión logística, SVM y RANDOM FOREST ofrecen los mejores resultados predictivos. El trabajo es relevante, pues el uso de los modelos dentro de un escenario pesimista permite aumentar la tasa de finalización de matrícula de seleccionados en un 1%, lo cual significa un beneficio directo de más de 30 millones de pesos solamente en concepto de pago de matrículas.

DEDICATORIA

“A tu dios y al mío”.

AGRADECIMIENTOS

Gracias a todos los que harán, hacen e hicieron posible este trabajo de título.

Al tata por las grandes enseñanzas que me diste a lo largo de tu vida.

A mis padres, por enseñarme que con cariño, tesón y trabajo todo es posible.

A mi amigo Klaus por las tantas historias y aprendizajes que tuvimos en la escuela.

A mi hermano por su apoyo incondicional en los momentos en que escribía.

A ti, por enseñarme que el amor es la fuente de energía más alta existente en el universo.

A todos los profesores que de algún modo transmiten el cuestionarse las cosas para entender los distintos fenómenos que nos rigen.

A todos aquellos que han contribuido en mi formación.

A Buchef, una facultad en que de formas misteriosas te conectas con la intuición.

"las carreras son de resistencia, no de velocidad"

Ruperto Ortiz Vargas

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE ILUSTRACIONES

1. INTRODUCCIÓN

Cada año cerca de 300.000 personas (entre egresados de 4to medio y egresados de otros años) se suscriben al proceso del sistema único de admisión universitaria chilena, donde tras rendir y obtener los puntajes de su prueba PSU, son seleccionados en base a sus resultados dentro de la casa de estudios en la carrera de su elección.

Dentro del sistema único de admisión universitaria (SUA) participan las 41 universidades adscritas al CRUCH y algunas universidades privadas, las cuales deben dar a conocer la oferta de sus carreras y puntajes de postulación al comienzo de cada nuevo proceso. Tras la selección de sus postulantes, todas estas universidades cuentan con los mismos periodos de matrícula.

La Universidad en estudio, corresponde a una de las universidades privadas más grandes del país. Si bien esta ha acaparado el mayor número de postulaciones a nivel nacional, tiene dificultades a la hora de efectivamente completar sus cupos, pues solamente cerca del 60% del total de alumnos (14.000 aproximadamente) que quedan seleccionados termina por finalizar su matrícula en el primer periodo establecido por el DEMRE.

El trabajo de tesis tiene como objetivo el producir una herramienta que permita predecir que convocados son poco propensos a matricularse, para así tomar acciones dentro del periodo de matrículas que permitan aumentar los chances que los seleccionados concreten su matrícula.

En la primera parte del trabajo se describe el proceso de admisión. A continuación, se presenta el estudio y procedimiento metodológico seguido de su realización. Finalmente son presentados los resultados de predicción obtenido por los modelos estudiados.

1.1 Antecedentes generales

1.1.2 Proceso de admisión universitario

El Departamento de Educación y Registro Educacional (DEMRE) es un organismo dependiente de la Universidad de Chile, cuya misión es desarrollar, aplicar y analizar la prueba de selección y admisión a la educación superior[1] dentro del Sistema Único de Admisión (SUA). Actualmente dentro de este proceso participan de forma nacional, simultánea y transparente las 27 universidades pertenecientes al consejo de rectores (CRUCH) y las 14 universidades privadas adscritas a este sistema.

Las universidades participantes (41) dentro del proceso de admisión se encuentran descritas a continuación (Tabla 1):

Tabla 1. Universidades adscritas al SUA.

Universidades del CRUCH	Universidades del CRUCH	Universidades Privadas Adscritas
Universidad de Chile	Universidad de Talca	Universidad Diego Portales
Pontificia Universidad Católica de Chile	Universidad de Atacama	Universidad Mayor
Universidad de Concepción	Universidad de Tarapacá	Universidad Finis Terrae
Pontificia Universidad Católica de Valparaíso	Universidad Arturo Prat	Universidad Andrés Bello
Universidad Técnica Federico Santa María	Universidad Metropolitana de Ciencias de la Educación	Universidad Adolfo Ibáñez
Universidad de Santiago de Chile	Universidad de Playa Ancha de Ciencias de la Educación	Universidad de los Andes
Universidad Austral de Chile	Universidad Tecnológica Metropolitana	Universidad del Desarrollo
Universidad Católica del Norte	Universidad de Los Lagos	Universidad Alberto Hurtado
Universidad de Valparaíso	Universidad Católica del Maule	Universidad Católica Cardenal Silva Henríquez
Universidad de Antofagasta	Universidad Católica de la Santísima Concepción	Universidad Autónoma de Chile
Universidad de La Serena	Universidad Católica de Temuco	Universidad San Sebastián
Universidad del Bío-Bío	Universidad de Aysén	Universidad Central de Chile
Universidad de la Frontera	Universidad de O´Higgins	Universidad Academia del Humanismo Cristiano
Universidad de Magallanes	----------	Universidad Bernardo O´Higgins

[1] http://www.uchile.cl/portal/presentacion/asuntos-academicos/demre/presentacion/110082/acerca-del-demre

El proceso de postulación tiene una duración de aproximadamente 10 meses desde la inscripción del postulante dentro del proceso hasta que se matricula en una universidad (Imagen 1). Para el proceso de admisión 2018 fueron 294.176 los inscritos a participar dentro del proceso.

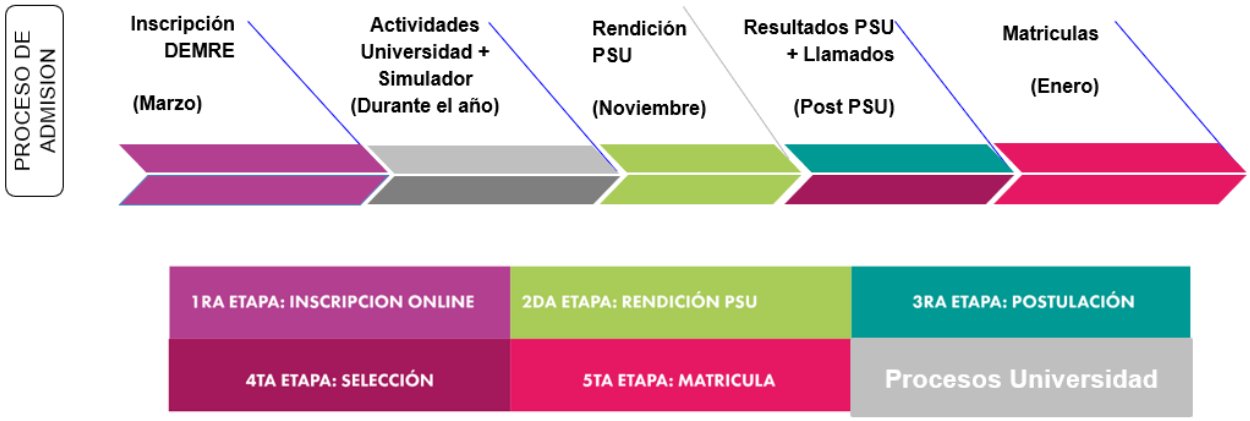


Imagen 1. Etapas admisión regular.

Es el consejo de rectores la entidad que establece las fechas, condiciones y plazos que se tienen dentro del proceso de cada año. Dentro de las etapas del proceso destacan los siguientes hitos:

1. **Inscripción al proceso de postulación:** durante esta etapa los estudiantes realizan la inscripción al proceso de inscripción del DEMRE, teniendo que ingresar información sociodemográfica de su familia (montos de ingreso familiar, cantidad de personas que trabajan, nivel educacional de los padres y hermanos, etc.) además de fuentes de financiamiento para futuros estudios (tanto en vivienda o becas). Dentro de esta etapa, para que el alumno pueda postular a cualquiera de las carreras de su elección, el postulante debe inscribirse y rendir las pruebas obligatorias de Lenguaje y Matemáticas, además de las pruebas de Ciencias sociales y/o Ciencias Naturales según corresponda.

2. **Rendición de la PSU:** dentro de esta etapa los estudiantes deben rendir las pruebas del proceso de postulación dentro de 2 días. En el primer día se rinden las pruebas de Lenguaje y Ciencias Naturales; y durante el segundo las pruebas de Matemáticas y Ciencias Sociales.

3. **Postulación**: transcurrido cerca de un mes desde la rendición de la PSU, los estudiantes conocen sus resultados de la PSU, además de las becas que tienen preasignadas por parte del Estado, dando inicio al proceso de postulación.

Esta etapa tiene un periodo de duración de 5 días para realizar su postulación a través del sitio del DEMRE, teniendo 10 opciones de carreras a las cuales postular (tomando un orden decreciente de prioridad). Dentro de esta etapa los postulantes demás de conocer los distintos criterios de postulación de sus puntajes a las distintas opciones tienen acceso a información de aranceles, duración de la carrera, mallas curriculares y campo laboral a futuro.

4. **Selección:** una vez publicados los resultados de postulación a las entidades de educación superior, se indica si el postulante fue convocado[2] o si quedó en lista de espera en las carreras a las cuales postuló sujeto a las vacantes de cada carrera.

5. **Matricula:** una vez publicados los resultados de la selección, se da paso al proceso de matrícula. En este los convocados tienen un plazo de 3 días para matricularse sin posibilidad de perder su cupo(Imagen 2).

   Es importante mencionar que el seleccionado al concretar su matrícula durante el primer periodo elimina automáticamente el resto de sus preferencias, teniendo el segundo periodo de matrículas ("lista de espera") para retractarse de su elección o cambiarse a otra carrera.

[2] Alumno que queda seleccionado en su primera carrera de preferencia.

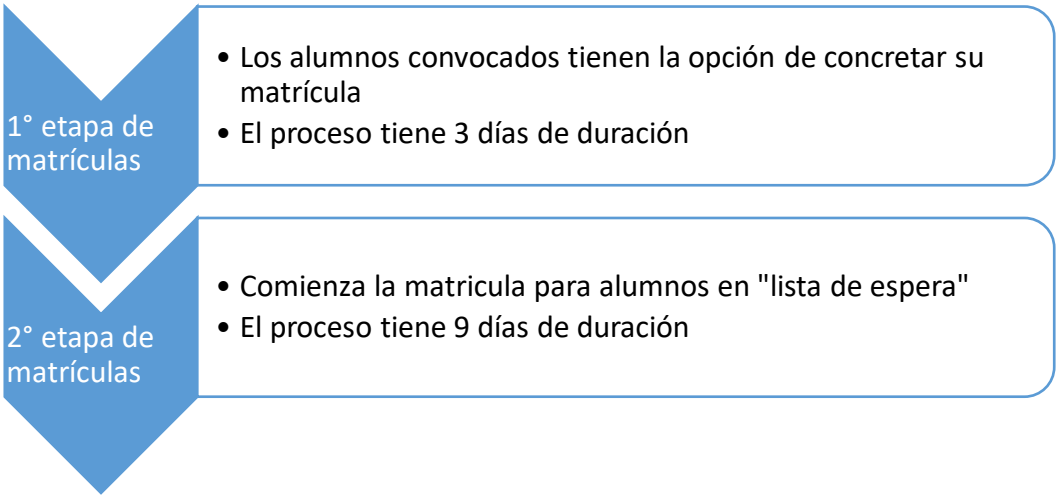


Imagen 2. Etapas proceso de matrículas.

La segunda etapa de matrículas tiene una duración de 9 días, donde los seleccionados en "lista de espera" pueden inscribirse dentro de alguna casa de estudios.

1.1.3 Información del proceso

Las universidades adscritas al Sistema único de Admisión disponen de la siguiente información a medida que avanza el proceso de selección universitaria:

1- Información de postulaciones: las universidades adscritas al proceso reciben información de todos los estudiantes inscritos en el proceso de admisión DEMRE; además de información con respecto a los colegios participantes como lo es la cantidad de matrícula, tipo de dependencia (municipal, subvencionado o privado), régimen educacional, GSE, etc.

2- Información con los resultados PSU: un día anterior a la publicación de resultados PSU, las universidades reciben los puntajes de los alumnos que rindieron la prueba, al igual que sus ponderaciones de puntajes. De forma adicional, se recibe información de las becas a las cuales postulo el alumno y el estado de preselección en que estas se encuentran.

3- Resultados de las postulaciones: un día antes de la publicación oficial de resultados, se reciben los resultados de las postulaciones a nivel de todas las universidades participantes del proceso, conociendo en detalle que alumnos son convocados a cada una de sus carreras.

   Esta información es conocida por la universidad un día antes del periodo de las matrículas.

4- Resultado de las matrículas: a medida que pasan los días de los periodos de matrícula, las universidades van ampliando su registro en cuanto a cuáles alumnos se terminaron matriculando en la respectiva casa de estudios, en qué carrera y facultad.

1.4 Características de la Organización en estudio.

La Universidad en estudio es una universidad privada chilena surgida en el año 1988. Actualmente ésta cuenta con distintas facultades en las ciudades de Santiago, Viña del Mar y Concepción, donde se ofrecen cerca de 60 carreras de pregrado y más de 50 programas de magister, logrando abarcar una matrícula de cerca de 49.000 alumnos (Tabla 2), siendo la universidad más grande a nivel nacional.

Tabla 2. Alumnos por universidades

Universidad	Categoría	N° de alumnos
Universidad en estudio	Privada	49.489
Universidad de Concepción	CRUCH	43.028
Universidad San Sebastián	Privada	40.287
Universidad Autónoma de Chile	Privada	34.840
Universidad de Chile	CRUCH	31.963
Universidad de Santiago de Chile	CRUCH	29.058
Universidad Diego Portales	Privada	23.073
Universidad Mayor	Privada	22.704
Universidad Austral de Chile	CRUCH	21.911
Universidad de Valparaíso	CRUCH	21.086

En el año 2012 como señal de compromiso con la transparencia y calidad en educación, la universidad entra a participar dentro del Sistema Único de Admisión (SUA), dándole reconocimiento al prestigio de la institución, por medio de contar con un proceso regulado y transparente de admisión, en la cual se ve obligado a dar a conocer la oferta de vacantes sus carreras, los puntajes de ingreso de forma anticipada y tener los mismos periodos de matrícula que las universidades adscritas.

De acuerdo con expertos en educación, esta universidad es considerada como de selectividad media (Meller, 2010), donde no se muestra un alto grado de selección, pero igualmente sus alumnos no son los que obtienen puntajes más bajos. El criterio de clasificación está en tener menos del 50% de sus alumnos con puntajes sobre 601, pero menos del 20% de ellos con puntajes bajo los 502 puntos. En definitiva, es una universidad que agrupa alumnos con puntajes entre 501 y 600 puntos.

2. JUSTIFICACIÓN DEL PROYECTO

Los procesos de admisión llevados a cabo durante los últimos años han mostrado que a pesar de los distintos esfuerzos que la universidad pone a lo largo del año en atraer futuros postulantes, no logra que la totalidad de los que son finalmente seleccionados finalice su matrícula.

Dentro del proceso de matrículas del proceso de admisión 2018, del total de seleccionados a matricularse dentro de la universidad en estudio, solamente el 54% (7.300 seleccionados) de estos concretó su matrícula dentro del primer periodo de matrículas (Imagen 3) dejando cerca de 6200 cupos sin completar. Situación similar es encontrada en proceso del año siguiente, en el cual cerca de un 39% de los cupos quedaron vacantes (5600 cupos aprox.).

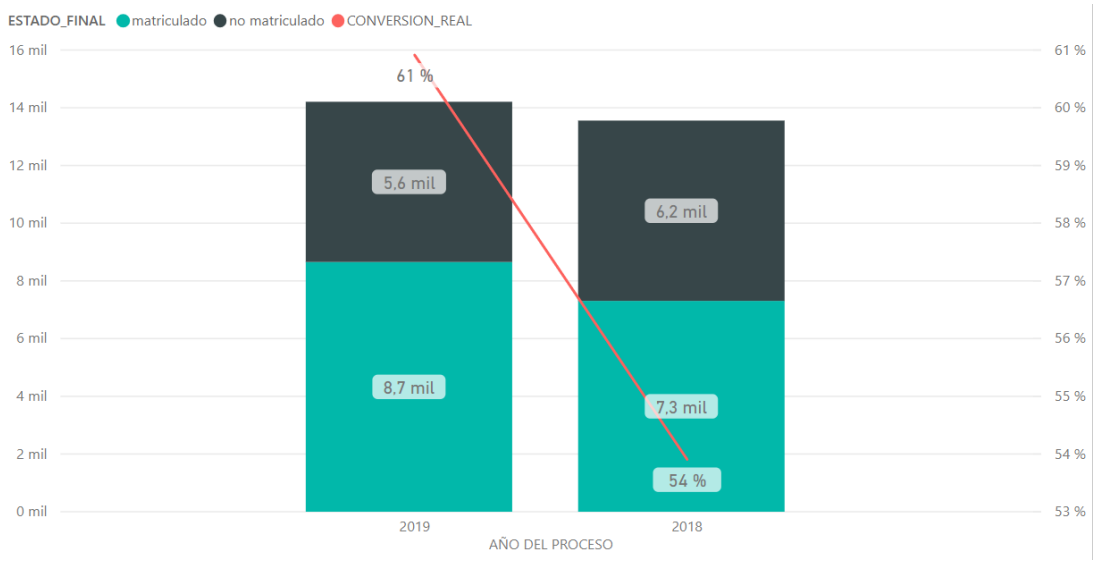


Imagen 3. Cantidad de convocados procesos 2018-2019.

Es importante destacar que durante el periodo de matrículas, las universidades participantes del proceso de admisión toman contacto con los postulantes a modo de incentivar que éstos se matriculen con ellos (vía beneficios de matrícula o arancel), lo cual impacta directamente en que parte de los seleccionados por la universidad en estudio, no concreten su matrícula. Una vez terminado el primer proceso de matrícula, se da paso a las matrículas de lista de espera, donde a modo de completar las vacantes disponibles se terminan otorgando becas que suelen ser mayores a las que se ofrecen dentro del primer proceso.

Dentro del proceso 2018, se puede apreciar que del total de seleccionados, cerca de un 24% terminó matriculándose en otra casa de estudios (Imagen 4).

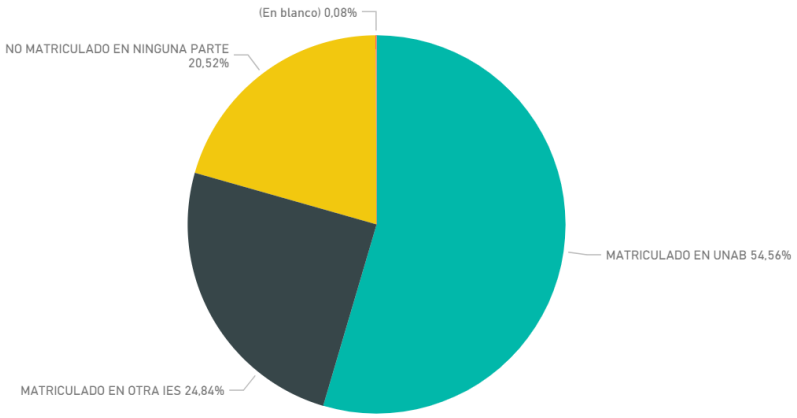


Imagen 4. Estado final de matrícula convocados 2018.

Este fenómeno sostenido a lo largo del tiempo, sumado al hecho que cerca del 50% de los seleccionados que no concretan su matrícula si lo hacen en otra universidad, permite plantear la hipótesis que el tener una herramienta que permita identificar los alumnos que no se matricularan, podría ayudar a aumentar la cantidad de matriculados por medio de realizar una mejor gestión final de contacto con el seleccionado durante el primer periodo de matrículas.

Tabla 3. Datos matrículas año 2018.

Matricula promedio	$408.171
Arancel promedio	$4.260.345
Alumnos matriculados	8004
Tasa actual matricula	54%
Beneficio actual	$37.366.802.064

El proyecto tiene relevancia, ya que el solo incremento de un 1% de la actual tasa de finalización de matrícula (Tabla 3 y 4), repercute en un ingreso de $32.670.007 solo por concepto de pago de matrículas, sin considerar arancel.

Tabla 4. Beneficios esperados del proyecto.

Incremento tasa de matricula	Beneficio extra matriculas	Beneficio extra arancel	Beneficio total
1%	**$32.670.007**	$340.998.014	$373.668.021
3%	$98.010.021	$1.022.994.041	$1.121.004.062
5%	$163.350.034	$1.704.990.069	$1.868.340.103
10%	$326.700.068	$3.409.980.138	$3.736.680.206

3. OBJETIVOS

Dentro de esta sección se detalla el objetivo general y los objetivos específicos del presente proyecto.

3.1 Objetivo general

Realizar un modelo de predicción que permita identificar los convocados con menor tendencia a matricularse dentro del primer periodo de matrículas, para realizar campañas de contacto que permita aumentar la actual tasa de conversión de matrículas.

3.2 Objetivos específicos

Para cumplir con el objetivo general, se detallan los siguientes objetivos específicos:

- Identificar los procesos y fuentes de datos ligados al proceso de selección universitaria y de contacto por parte de la universidad con sus postulantes convocados.
- Investigar los factores que hacen que el convocados no finalice su matrícula.
- Determinar las variables que poseen mayor incidencia dentro de la opción de finalizar la matricula dentro de la universidad en estudio.
- Generar un modelo que permita identificar a los convocados con menor tendencia a matricularse.
- Evaluar económicamente los resultados del modelo dentro de un caso de uso.

3.3 Alcances

1. Los datos a utilizar serán los provistos por la contraparte interesada en el proyecto, sin considerar mediciones en terreno o uso de otras fuentes de información.
2. Al trascurso de la memoria se diseñará un prototipo funcional del modelo propuesto, sin considerar la implementación de este.

3.4 Resultados esperados

1. Identificar fuentes de datos que se recolectan desde la inscripción de los alumnos desde el inicio del proceso DEMRE, hasta que se matriculan en la universidad.
2. Comprender el funcionamiento del proceso de matrículas.
3. Obtener las variables que tengan relevancia dentro de la decisión de los convocados a finalizar su matrícula.
4. Desarrollar un modelo de predicción que estime la tendencia de que el alumno finalice su matrícula.
5. Con los resultados se pretende entregar recomendaciones sobre formas de aumentar la tasa de conversión de matrícula.

4. MARCO CONCEPTUAL

Este capítulo presenta el conjunto de conceptos y teorías sobre el cual se estructura el desarrollo del trabajo de título, los cuales surgen de la revisión de los métodos utilizados en diversos trabajos de marketing cuantitativo y trabajos de Datamining que se apoyan principalmente en el uso de distintas fuentes de información para la extracción de conocimiento útil a un negocio especifico a través de la metodología del KDD (*Knowledge Discovery in Databases*).

4.1 Conceptos asociados al proceso de admisión universitario

Dentro del proceso de las matrículas, es importante destacar las diferencias entre los distintos tipos de alumnos que pudiesen existir dentro de la selección, destacándose 2 clases:

i) Alumno Convocado:

Se define como el alumno que posteriormente al proceso de selección, queda llamado a poder matricularse dentro del primer periodo de matrículas.

ii) Alumno en Lista de espera:

Corresponde al alumno que posteriormente al proceso de selección, no queda llamado al primer periodo de matrícula, pero si puede concretar su opción en la segunda etapa de matrículas en caso de que se den las condiciones de cupos disponibles.

iii) Nivel de conversión:

Indicador en base al cual se realizan la comparación entre la cantidad de alumnos que son llamados a matricularse a la universidad, versus la cantidad que efectivamente concreta su matrícula.

$$\boldsymbol{Conversion = \frac{Total\ matriculados}{\text{Total convocados}}}$$

4.2 Revisión bibliográfica

Un entendimiento adecuado del actual contexto de la educación da cuenta con hechos empíricos como ha ido evolucionando la matrícula universitaria a partir del año 2000, donde se alcanzaba 151.538 estudiantes, en contraste al año 2010 en que la cifra se eleva a 562.583 universitarios, exhibiendo mayor inclusión de los deciles más bajos dentro de la educación universitaria debido a un mayor acceso de financiamiento vía becas o créditos (Meller, 2011).

El aumento de la cantidad de alumnos dentro del sistema universitario ha permitido la realización de distintos trabajos de titulo relacionados con la universidad en estudio, investigando factores relacionados con como compite la universidad con sus pares, cómo realizar difusión en colegios y cómo asignar becas de forma más eficiente. Dentro de estos trabajos se destacan aportes los (Núñez, 2013) en donde al identificar la competencia de distintas carreras y universidades, realiza un análisis en el cual concluye como elementos diferenciadores en la cantidad de postulaciones a la universidad elementos como los *puntajes PSU, existencia o no de contacto previo con la universidad, además del conocimiento de la existencia de becas preasignadas*. Con relación al contacto que pueden tener los alumnos con la universidad, es que (Espinace, 2014) realiza un mix de acciones de difusión con las cuales la universidad debiese actuar sobre distintos colegios para impactar en un mayor número de postulaciones. En la misma línea, (Nuñez, 2018) investiga sobre la efectividad de las acciones, recomendando un número máximo de acciones a realizar por semestre, encontrando cuales acciones tienen un mayor impacto según el tipo de colegio.

Dentro de los trabajos que involucran como mejorar el otorgamiento de becas en los distintos programas de estudios ofrecidos, (Barrientos, 2014) realiza un modelo para estimar la correcta asignación de recursos que cada una de las carreras debiese tener en el simulador de beneficios, *destacando que el poseer o no beneficios hace una gran diferencia en que un alumno se matricule o no dentro de la universidad.*

4.3 El proceso KDD

El proceso del KDD (*Knowledge Discovery in Databases*) se refiere al proceso de adquirir conocimiento útil para la gestión de un negocio, mediante la búsqueda de patrones y relaciones que pueden resultar inesperadas dentro de sus bases de datos. Este consiste en un proceso iterativo de cinco etapas (Barrientos & Rios, 2013):

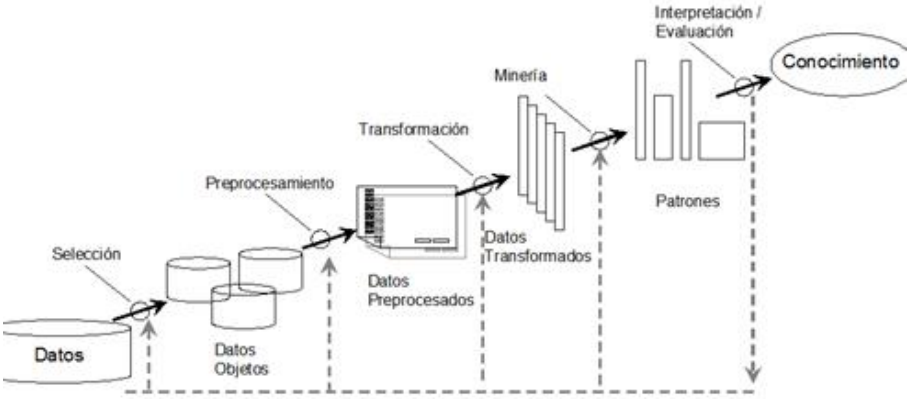


Imagen 5. Proceso KDD.

i) Integración o selección: etapa en la cual se escogen las variables y fuentes de datos a ser consideradas en el proceso completo, por lo que se refiere a la creación de la base de datos con la cual se trabajará.
ii) Preprocesamiento: dentro de esta sección se procede al análisis y limpieza de los datos, realizando el tratamiento correspondiente en caso de valores ausentes o fuera de rango, empleando diversas técnicas de imputación de valores.
iii) Transformación: acá se generan nuevas variables a partir del estudio de la naturaleza de las variables originales.
iv) Minería de datos: esta etapa consiste en la aplicación de análisis de datos para descubrir un algoritmo ad-hoc que produzca una enumeración de patrones a partir de los datos y que los produzca tomando en cuenta la capacidad computacional disponible.
v) Interpretación y evaluación: en esta última etapa se toman medidas de evaluación de los resultados técnicos y comerciales, de manera que los modelo sirvan para tomar acciones que solucionen el fenómeno estudiado.

4.3.1 Modelos de Datamining

Dentro de la etapa de minería de datos, se encuentra una gran variedad de modelos y perspectivas a aplicar, dentro de los que se encuentran diversos algoritmos. Se procede a describir los más usados dentro de la investigación:

Regresión: este tipo de modelo consiste en el estudio de dependencia de una variable dependiente, respecto a un conjunto de variables explicativas, con el objetivo de estimar o predecir la media de la primera variable en términos de los valores conocidos. Este modelo sirve para predecir y clasificar.

En el caso de la clasificación, la regresión no resulta muy efectiva, puesto que la variable a predecir es de naturaleza nominal, sin embargo, las regresiones logísticas solucionan este problema.

SVM: Algoritmo que a través de hiperplanos crea una clasificación lineal de clases. En general los SVM usan un mecanismo llamado kernel, que es esencialmente la distancia entre dos medidas. El algoritmo encuentra un "borde" que maximiza la distancia entre los miembros de distintas clases.

Naive Bayes: Algoritmo basado en probabilidades condicionales y conteo de frecuencias, lo cual lo hace un modelo que no entiende de la interdependencia que pueda existir entre las variables. En el análisis bayesiano, la clasificación final es producida como una combinación de la información probabilística y de máxima verosimilitud.

Entre sus ventajas principales esta su fácil implementación y que puede ser usada para multi-clasificación. La principal desventaja es que si se le pasa una "observación" no vista antes, la asigna automáticamente como el caso por default.

Árboles de decisión: modelo predictivo que a partir de un conjunto de datos elabora diagramas de reglas que permiten representar o categorizar un fenómeno, usualmente estos modelos se presentan en forma de grafos. Es un modelo que puede ser utilizado para representar tanto modelos regresivos como de clasificación.

De forma general, el árbol de decisión consiste en un grafo donde existe un nodo único o parental, el cual contiene instancias a tener en cuenta en el modelo. Es un modelo de fácil implementación e interpretación, pero pueden llegar a crear interacciones muy complejas entre los datos.

Dentro de sus distintas variantes podemos encontrar:

Random Forest: algoritmo supervisado, variante de los árboles de decisión, en los cuales como indica su nombre, crea un conjunto de árboles de decisión(bosque) tomando variables del set de forma aleatoria para ir armando los árboles. Es un algoritmo flexible a utilizar en tareas de clasificación o predicción.

4.3.2 Métricas de evaluación

Las medidas de evaluación de modelos generalmente usadas se basan en el uso de la tabla de contingencia que describe las instancias predichas acertadas y erróneas [Imagen 4]. También conocida como matriz de confusión, contiene información acerca de las clasificaciones reales y las predichas por el sistema de clasificación. El esquema de ésta para un caso de clasificación binaria es la siguiente:

Tabla 5. Tabla de contingencia.

		Clase predicha	
		Clase = 1	Clase = 0
Clase actual	Clase = 1	TP (true positive)	FN (false negative)
	Clase = 0	FP (false positive)	TN (true negative)

A partir de la tabla de contingencia se definen las siguientes métricas de evaluación:

i) **Accuracy**: métrica que corresponde al ratio entre los casos correctamente clasificados, por sobre el total de predicciones.

$$\textbf{Accuracy} = \frac{(\textbf{TP} + \textbf{TN})}{(\textbf{TP} + \textbf{FP} + \textbf{TN} + \textbf{FN})}$$

Si bien es una métrica que indica que tan bien el modelo predice las clases, es importante notar que puede presentar problemas de interpretación en casos donde las clases posibles a predecir están balanceadas, por lo cual, dependiendo del caso,

es importante revisar en cuál de las clases se está interesado en acertarle más. Para los fines anteriores es útil revisar las siguientes métricas:

ii) ***Sensitivity***: métrica que consiste en ratio de los casos positivos correctamente clasificados entre todos los casos positivos.

$$\boldsymbol{Sensitivity} = \frac{\mathbf{TP}}{(\mathbf{TP} + \mathbf{FN})}$$

iii) ***Specificity***: métrica que corresponde al ratio de los casos positivos correctamente clasificados entre todos los casos clasificados como positivos.

$$\boldsymbol{Specificity} = \frac{\mathbf{TP}}{(\mathbf{TP} + \mathbf{FP})}$$

iv) **AUC:** métrica utilizada en problemas de clasificación donde se vienen variados puntos de corte para una probabilidad predicha. Esta corresponde al área bajo la curva ROC, otorgando información de que tan bien el modelo es capaz de distinguir entre dos clases. La curva ROC se visualiza como el ratio entre TPR (True positive rate) y el FPR (false positive rate)

True positive rate: TP/(TP + FN)

False positive rate: FP / (FP + TN)

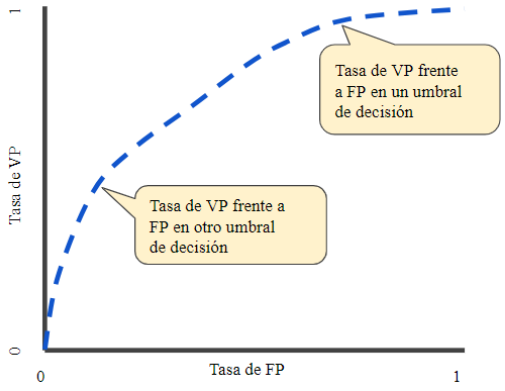


Imagen 6. Curva ROC.

5. METODOLOGÍA

Esta metodología de trabajo toma acción y protagonismo luego de haberse definido junto al cliente interesado el diagnóstico inicial y problemáticas relacionadas al fenómeno de las matrículas. Por lo tanto, expone todo lo que se realizó después de haber definido el proyecto a trabajar.

La columna vertebral de la metodología aplicada se basa principalmente en las etapas del KDD para la creación de los modelos. A continuación, se describe la metodología que se utilizará dentro de este trabajo para dar cumplimiento a los objetivos planteados, usando los datos de los procesos de admisión 2018 y 2019.

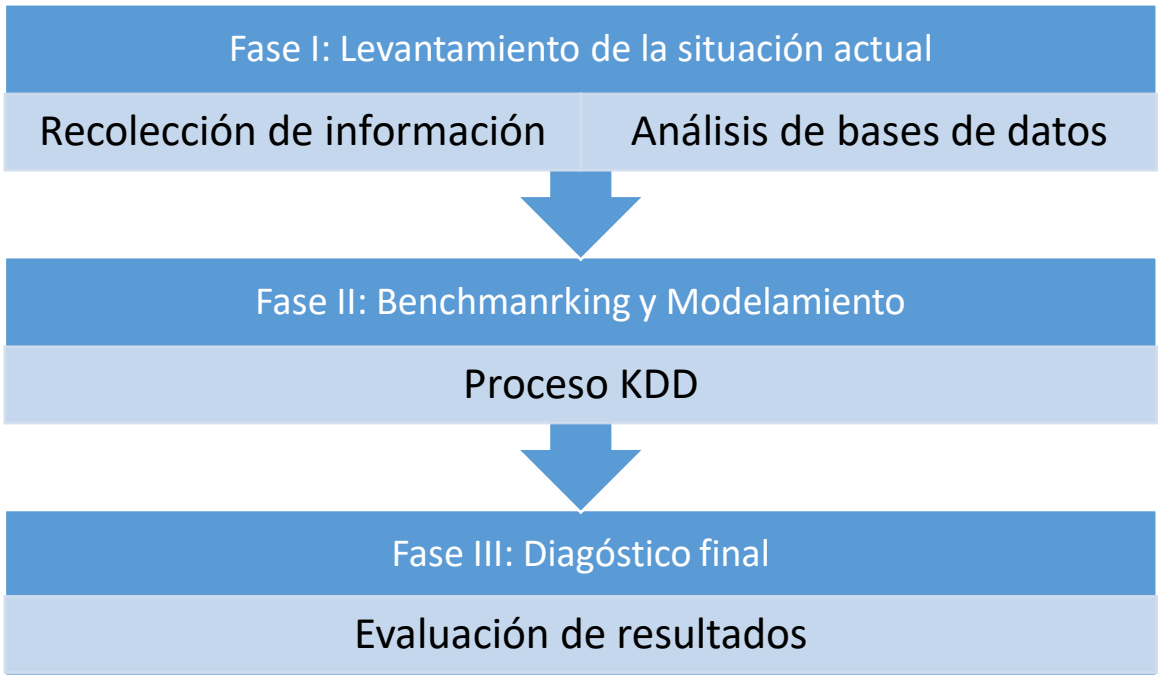


Imagen 7. Esquema de la metodología.

5.1 Levantamiento de la situación actual

Los principales focos de información que se buscaron recabar en esta memoria descansan en dos ejes fundamentales: (1) Conocer el proceso de admisión universitario, integrando información relevante declarada por la contraparte interesada para hacer un correcto diagnostico con la información de los procesos 2018 y 2019 (2) identificar las distintas fuentes de información ligadas al proceso de admisión y a la interacción que la universidad realiza con sus futuros postulantes.

Información de los solicitantes

Se realizan entrevistas semiestructuradas con los principales involucrados dentro del proceso de monitoreo de matrículas, para interiorizarse de las fechas de los distintos procesos, vías de contacto en que la universidad se comunica con sus postulantes. Como información base se tiene:

- Información proveniente del DEMRE.
- Opinión del performance del funcionamiento del proceso de matrículas en base a experiencia.
- Cantidad de convocados y matriculados dentro y fuera de la universidad en estudio.
- Conocimiento sobre el proceso de monitoreo de matrículas.
- Conocimiento de las fuentes de datos a utilizar.

El levantamiento de información comprende principalmente el actual funcionamiento del sistema de admisión y en específico del proceso de matrículas, con todos los hitos importantes de los cuales se pueden obtener datos, con los cuales obtener un panorama general de la complejidad del proceso en cuanto a las distintas carreas, tipos de convocados, causas de no matricula, entre otros. Además, se destaca la importancia de la revisión de temas de trabajo anteriormente realizados con la universidad en estudio, se modo de entregar una solución que utilice las herramientas y métodos usados actualmente por la industria.

5.2 KDD

La información recolectada proviene de las bases de datos entregadas por el DEMRE en las distintas etapas del proceso de admisión, así como también de las bases de datos relacionadas con los principales procesos de difusión y captación de alumnos que hace la entidad participante dentro del proyecto desde el comienzo del año hasta el periodo de matrículas. La realización del análisis exploratorio de la data consolidada tomando el nivel de conversión de matrícula como indicador de gestión permite intuir la importancia relativa de algunas variables a usar en los modelos.

5.2.1 Adquisición, integración y selección de datos

La adquisición de datos se realizará utilizando el lenguaje de SQL directamente desde el gestor de bases de datos de la entidad patrocinante del proyecto. La consolidación de la información de las bases de datos disponibles se realiza con el objetivo de generar un set de información que permita identificar las variables que tengan incidencia en identificar a los alumnos con posibilidad de matricularse.

Se utiliza el uso del lenguaje de programación R para la consolidación y limpieza de datos. Se recomienda esta herramienta principalmente por ser open-source, además de dar la flexibilidad de que esta investigación pueda ser fácilmente reproducida en el futuro.

El inicio de la consolidación de datos comienza con el uso de la base de datos de convocados, de la que se rescatan los campos "*RUT*" y "*Código de la carrera*" para ir agregando información del convocado, como información socioeconómica, puntajes PSU y participación en distintas actividades relacionadas con la universidad.

5.2.2 Preprocesamiento de datos

Con la base ya consolidada se realiza un análisis exploratorio para ver la consistencia de la información en caso de que existan datos erróneos, como valores que se encuentren fuera de rango(*outliers*), mal imputados o nulos. Se hace un análisis descriptivo de las variables disponibles, además de visualizar su distribución empírica de ser posible. Existen

diferencias en la meta-data de alguno de los valores, donde por ejemplo, existen variables numéricas que se guardan como texto.

En general, hay que destacar la buena integridad de la información con respecto a la información socioeconómica declarada.

5.2.3 Transformación de datos

Para efectos del modelamiento futuro, se creó una variable binaria para representar la variable de interés (se matricula o no).

Se generan variables binarias para captar distintas características de los alumnos convocados relacionadas a su colegio de origen, beneficios que tiene asignados o si ha tenido contacto con la universidad. Un ejemplo de lo anterior se puede ver dentro de la variable "tipo de colegio de procedencia" del convocado (1: Particular, 2: Subvencionado, 3: Municipal) la cual es transformada a 3 variables binarias que indican si pertenece o no a la nueva categoría creada. De igual forma, la información de actividades como si el alumno realizó simulación de beneficios dentro del el simulador web de beneficios de la entidad es transformada a variable binaria, indicando con un 1 si el alumno simuló y 0 el caso contrario.

Variables continúas relacionadas a distancias y montos de dinero son reformuladas haciendo uso de la transformación logarítmica. Las variables con gran cantidad de valores como "decil" del postulante o de los puntajes obtenidos, son categorizadas en distintas cantidades de tramos.

5.3.4 Modelamiento

Se realiza un contraste entre las distintas variables recolectada y el indicador de conversión para notar a priori cuales pudiesen tener mayor incidencia. Tomando un set inicial de **16** variables. Se procede a la fase de prueba de distintos modelos supervisados con el fin de conocer patrones previamente desconocidos dentro de los datos.

La base de datos será dividida en un set de entrenamiento[3] y testeo, de modo de evitar que los modelos a calibrar queden sobre ajustados y no posean poder predictivo. Como datos de entramiento. Se utiliza la información del proceso 2018 para entrenar y se testea y selecciona el mejor modelo en base a la información del proceso del 2019.

Como modelos testeados están: Random forest, regresión logística, Naive Bayes y super vector machine.

5.3.5 Interpretación y evaluación de los resultados

Dentro de esta etapa se mide la capacidad predictiva de los modelos en base a su poder predictivo al usar como set de testeo la infromacion del año 2019, realizando una comparación usando distintas metricas para seleccionar el modelo que responda de mejor manera a los objetivos de esta investigación.

Se testean las soluciones entregadas por los distintos modelos, en base a las metricas de *Acurracy, Sensitibity, Specificity, PosPred y NegPred.*

Cada tipo de modelo es comparado en su versión global(que no distingue entre tipo de convocados) contra una en la cual se calibran distintos modelos por tipo de convocado y se ensamblan sus resultados (para ver si existen mejoras en el nivel predictivo, discriminando principalmente por la métrica de accuracy y pospred).

El testeo de resultados se realiza comparando los output de los modelos dentro de distintos escenarios simulados de uso en que se ofrece beca de arancel a los convocados identificados como no matriculados, variando las tasas de respuesta positiva que se pudiesen tener por parte del total de los contactados entre un 1% y 35%.

La evaluación economica de los distintos escenarios simulados de contacto, exibe los beneficios y costos(minimo, maximo y promedio) estimados a partir de 100 simulaciones que se obtiene con cada uno de los modelos de machine learning utilizados, ademas de contrastar estos resultados con las respuestas que se obtendrian de un modelo aleatorio y de otro que permita predecir con un 100% de efectividad.

[3] Se excluyen observaciones de las carreras “Alta conversión”.

6. DESARROLLO

Dentro de esta sección se explican los principales pasos llevados a cabo a partir de la metodología propuesta.

6.1. Obtención y selección de los datos del proceso de admisión

Para la realización de este trabajo se han utilizado como fuentes de información bases de datos correspondientes a los procesos de admisión universitaria 2018 y 2019, además de datos internos sobre campañas de actividades y contacto de alumnos por parte de la universidad en estudio. Dentro de la tabla 6 se detallan las bases de datos utilizadas.

Tabla 6. Bases de datos utilizadas.

Bases de datos	Descripción
Inscritos DEMRE (295.536 observaciones)	Información de los alumnos inscritos dentro del proceso de selección universitaria.
Resultados PSU (266.221 observaciones)	Información con respecto a los resultados obtenidos por el alumno dentro de las distintas pruebas de selección universitaria.
Actividades (223.869 observaciones)	Información de actividades promocionadas por la universidad, dentro de las que participan los alumnos.
Simulaciones (154.442 observaciones)	Información obtenida del simulador de becas y beneficios de la universidad.
Llamados (21.426 observaciones)	Información de alumnos contactados por el call-center dentro de los días de conocimiento de los resultados de la PSU.
Postulaciones (26.686 observaciones)	Información de los alumnos postulantes dentro de los procesos una vez conocidos sus puntajes.
Matriculas (9.849 observaciones)	Información de los alumnos matriculados dentro del proceso de admisión, donde se destacan campos como carrera elegida, puntajes, colegio.

Como principales consideraciones dentro de los datos destacan:

1- Se toma en consideración la información de los procesos realizados en los años 2018 y 2019.
2- Selección solamente de los alumnos convocados dentro del primer periodo de matrículas.
3- Un diseño muestral representativo para el país y sus 16 regiones.

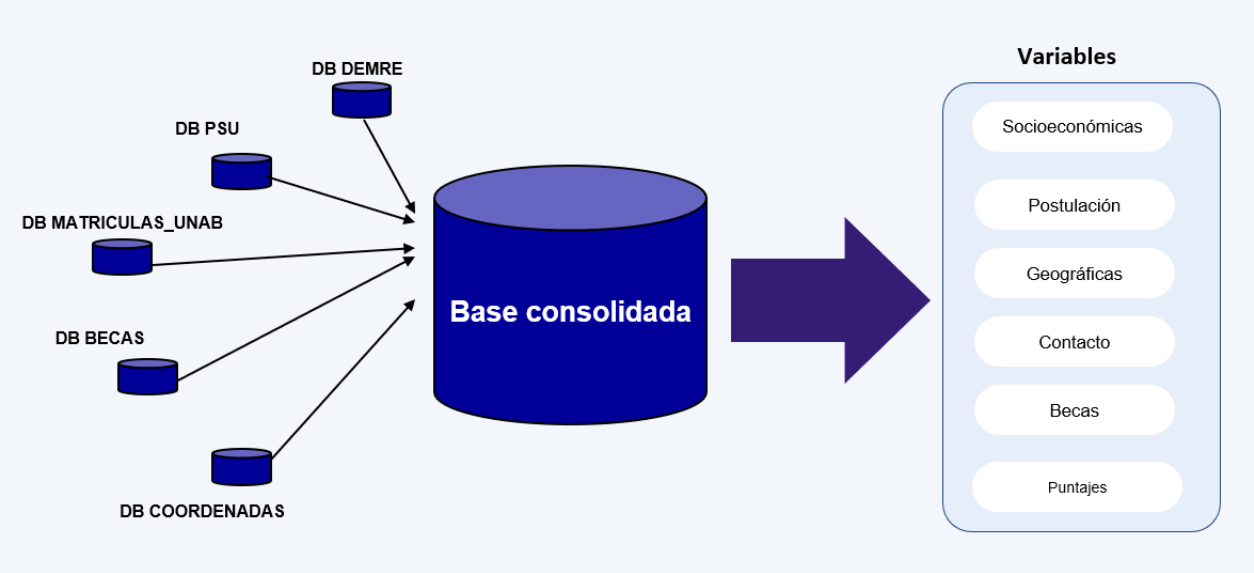


Imagen 8.Consolidación de bases de datos.

Finalmente se consolida una base de datos que contiene 46.955 observaciones que contiene información de todos los postulantes de los procesos 2018 y 2019 dentro proceso de matrículas.

6.2 Análisis descriptivo del proceso de matriculas

Dentro de esta sección se obtiene la primera mirada a las cifras del proceso de matrículas considerando a los alumnos convocados en los procesos de admisión 2018 y 2019. En general se nota que la mayor cantidad de postulaciones está concentrada dentro de la región metropolitana (75.3%), seguida por Viña del Mar (24.7%) y Concepción (20.5%) (Imagen 9). En cuanto a las tasas de conversión de matrícula para los procesos 2018-2019 se encuentran entre un 54% y 60% respectivamente.

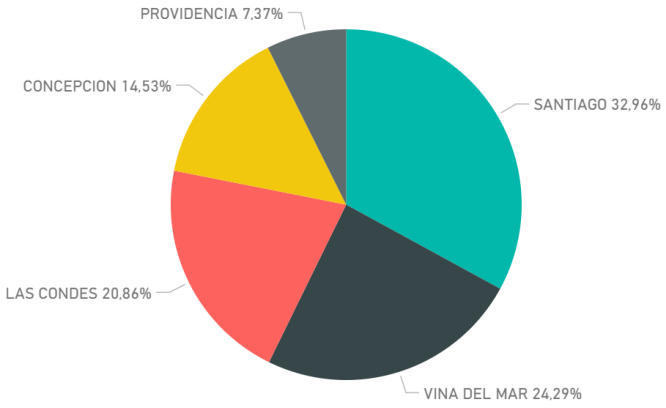


Imagen 9.Cantidad de postulaciones por comuna del campus (proceso 2018).

Durante los días de formalización de matrículas, se puede observar que cerca del 60% de los matriculados totales lo hace durante el primer día, con ritmos decrecientes en días posteriores, mostrando lo importante que pudiese ser un contacto anticipado con los seleccionados a modo de aumentar la cantidad de matriculados.

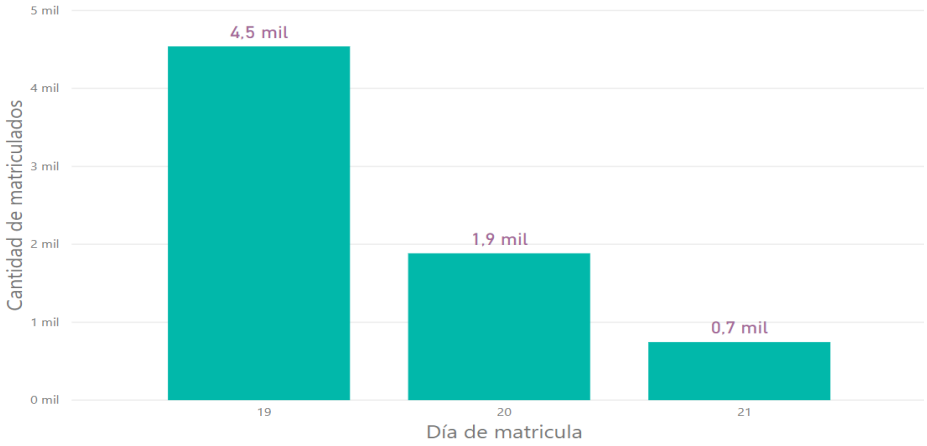


Imagen 10.Matriculados por día (proceso 2018).

6.2.1 Análisis de las carreras y sus convocados

Al analizar las distintas carreras tomando como variables de comparación el valor de sus aranceles, puntajes de ingreso y cantidad de convocados (representado por el tamaño de los círculos) (Imagen 11), se puede observar que las carreras ligadas a la salud (Medicina y Odontología) resaltan notoriamente del conjunto total por tener los mayores Aranceles y Puntajes de ingreso.

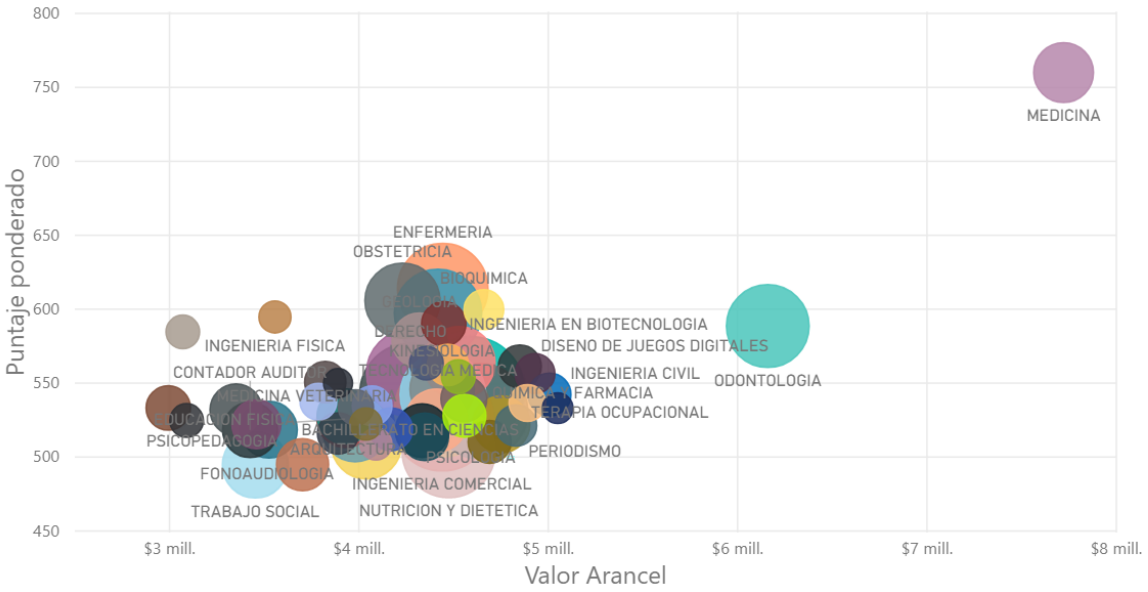


Imagen 11.Comparación aranceles contra puntajes ponderados promedio.

Al contrastar el nivel de conversión de las carreras con los puntajes ponderados, se nota que las carreras al área de la salud y las ingenierías poseen las carreras con mejores niveles de conversión (entre 53 y 78% del total de conversión). Por otra parte, carreras de facultades como humanidades, educación, ciencias biológicas y arquitectura, arte y diseño) como las carreras en mayoría con menores niveles de conversión de matrícula. Dentro de esta clasificación se nota que Medicina y Odontología vuelven a destacar por tener convocados con puntajes mucho mayores al conjunto total con 588 y 759 puntos respectivamente.

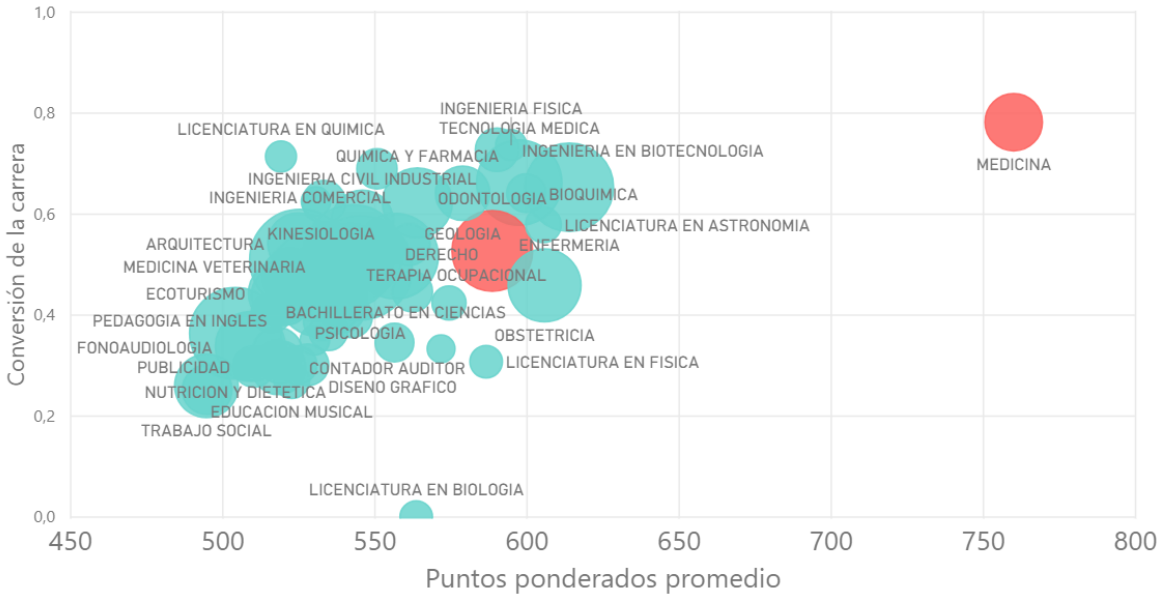


Imagen 12.Comparación de aranceles contra conversión del programa.

Al realizar la caracterización anterior de las carreras tomando en consideración la cantidad de convocados, porcentaje de conversión de la carrera, puntos ponderados y valor de los aranceles, se nota la existencia de 6 carreras que se dejan fuera de los análisis por tratarse de carreras con muy poca dificultad para llenar sus cupos (Carreras de alta demanda). En general, se puede intuir que un alumno que queda convocado a una carrera de alta conversión con puntajes altos tiene mejores chances que terminar matriculándose, en contraposición con un convocado con altos puntajes en una carrera de baja conversión.

6.2.2 Exploración de variables de interés

Se realiza la inspección visual de los datos correspondientes a los convocados dentro del primer periodo de matrícula, en contraste al indicador de nivel de conversión entendido como:

$$Conversión = \frac{Total\ de\ matriculados\ segun\ la\ variable\ i}{Total\ de\ alumnos\ convocados\ segun\ la\ variable\ i}$$

Con respecto a la elección de preferencia de la carrera al momento de la postulación, se puede notar que los alumnos seleccionados que se matriculan tienden mayoritariamente a poner las carreras de la universidad dentro de sus primeras 3 opciones de postulación (Imagen 13), teniendo índices de conversión que van desde un 73% para los que ponen la universidad en estudio en su primera opción, hasta un 39% para los que la ponen en opción 3.

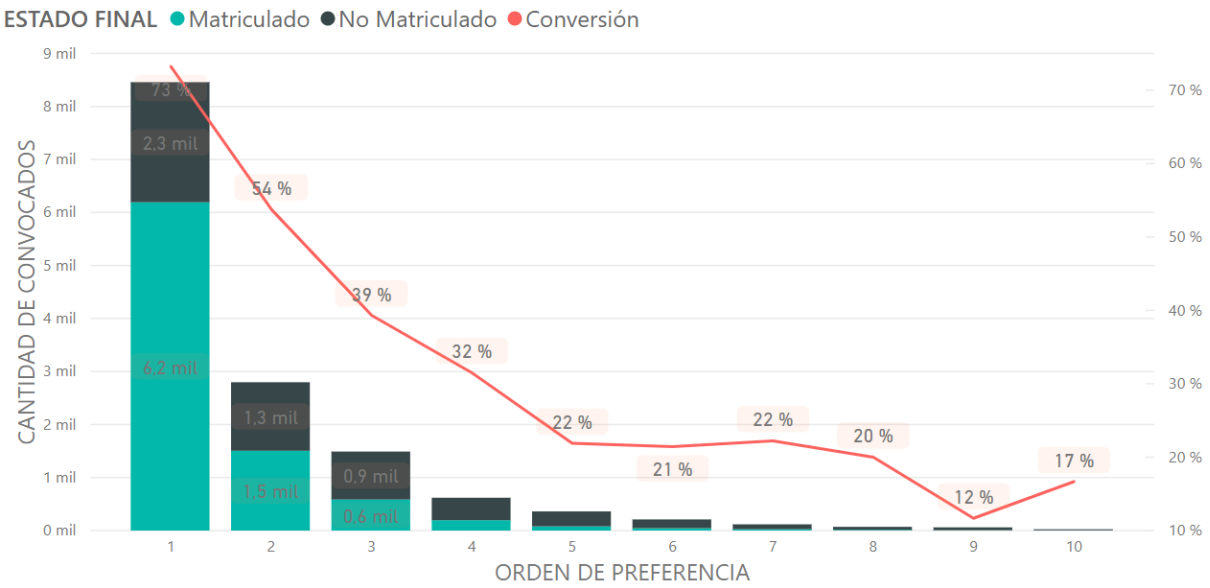


Imagen 13.Preferencia de postulación de alumnos convocados.

.

Al analizar a los convocados según el tipo de dependencia de su colegio, se puede observar una alta presencia de colegios particulares subvencionados (Imagen 14). Al revisar los niveles de conversión, se ve empíricamente que los colegios particulares representan el mayor nivel de conversión (74%), seguidos por los colegios subvencionados (60%) y colegios municipales (49%).

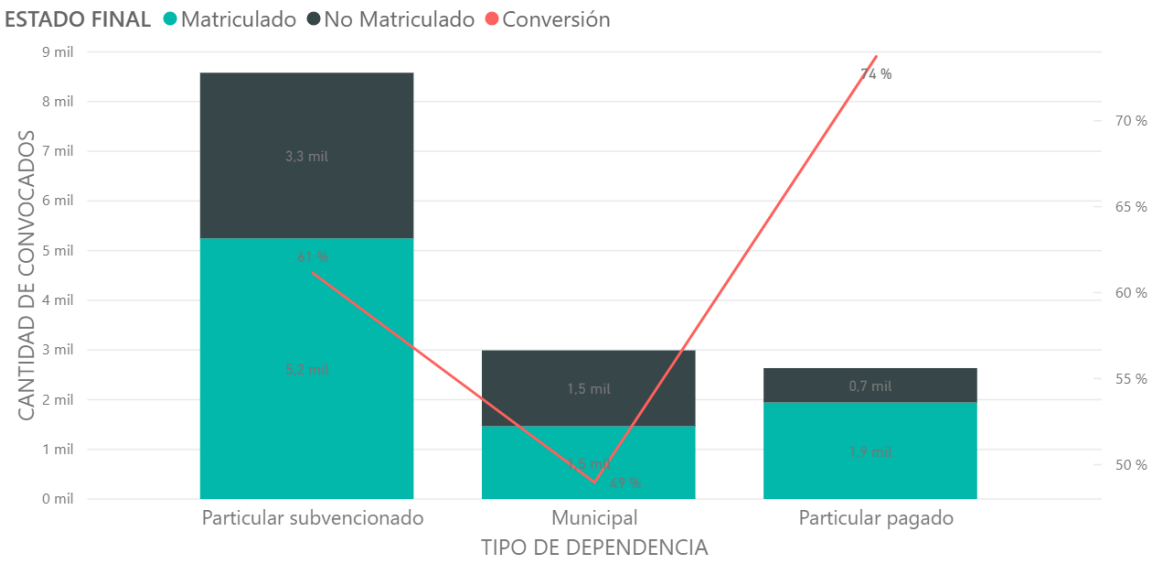


Imagen 14. Alumnos convocados según distintos tipos de colegio.

En cuanto al año de egreso de los postulantes, se nota que en su mayoría (más del 90%) corresponde a alumnos de hasta 3 años de diferencia entre que egresan del colegio y postulan al proceso de selección universitaria.

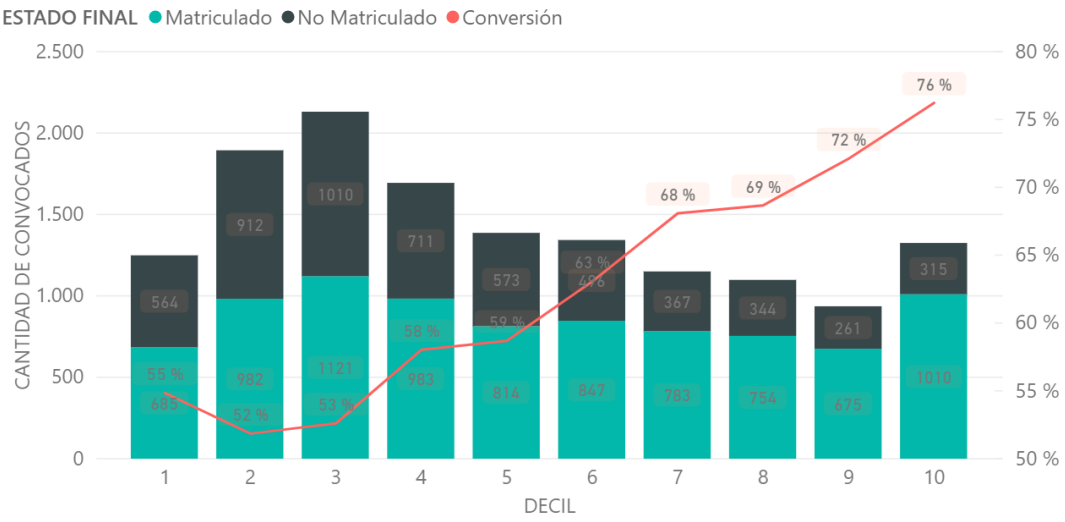


Imagen 15. Cantidad de convocados matriculados según Decil.

Al notar las diferencias de los convocados por su nivel socioeconómico, se puede notar que a medida que se avanza en los deciles, existe una tendencia al alza del nivel de conversión, fluctuando dentro de un 55% para decil 1; y 76% para el Decil 10(Ingresos superiores a los $600.000) (Imagen 15).

Dentro del contexto universitario chileno, se ha destacado por ser reconocido como los sistemas universitarios más caros del mundo, con la particularidad de además que la gente se endeuda entre cerca del 50% de sus ingresos, lo cual hace evidente el acceso a becas y créditos sea en muchos casos la única opción para seguir con estudios superiores.

Al analizar las variables relacionadas con distintos tipos de becas que tienen los convocados, se nota que el obtener la beca de excelencia académica que postulan son una proporción demasiado baja en relación al resto de la muestra, en donde el ratio entre convocados y matriculados es aproximadamente 0.5 (Imagen 16), sin embargo, el estar preseleccionado con CAE(Imagen 17), estar becado por becas del Gobierno (Imagen 18) o becas de la universidad (Imagen 19) parecen ser más relevantes con niveles de conversión del 66%, 68% y 72% respectivamente.

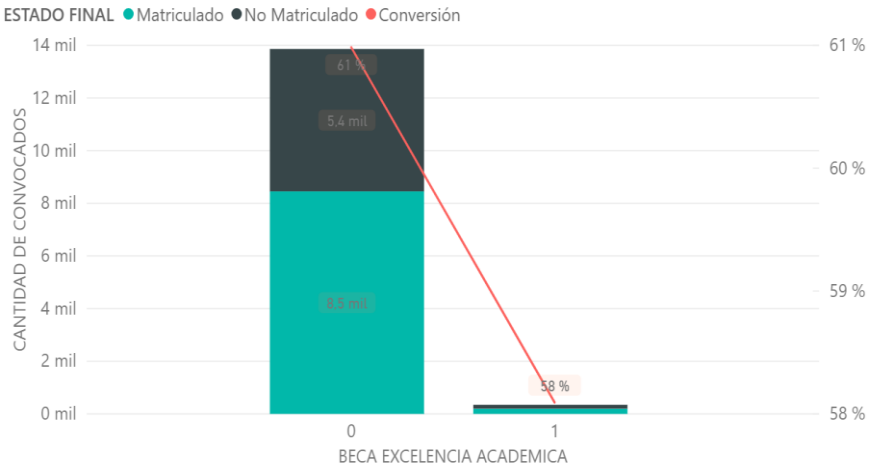


Imagen 16. Cantidad de matriculados con BEA.

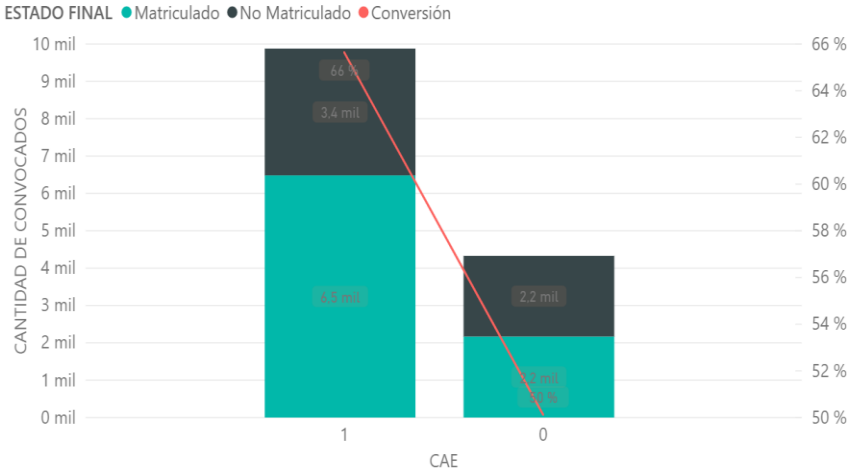


Imagen 17. Cantidad de matriculados con CAE.

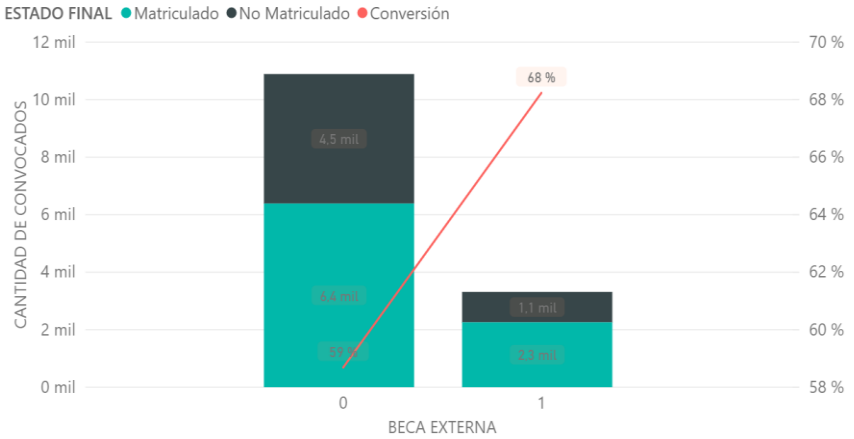


Imagen 18. Cantidad de matriculados con BECA EXTERNA.

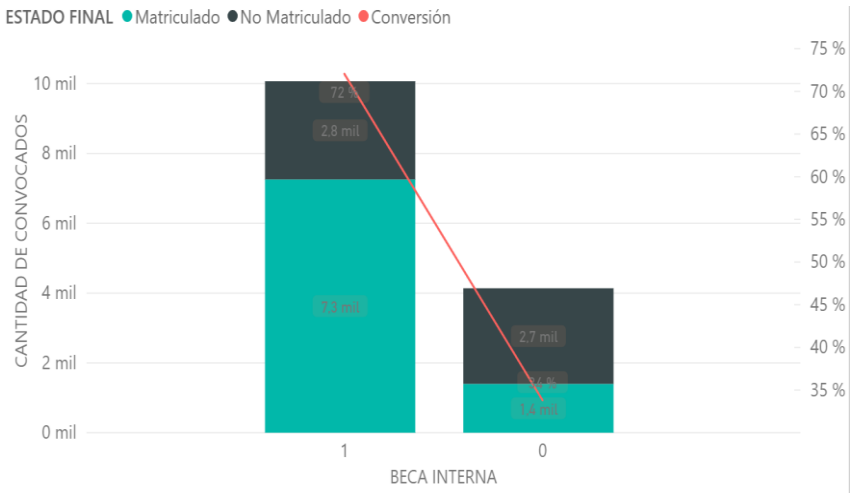


Imagen 19. Cantidad de matriculados con BECA INTERNA.

Dentro de la imagen 20, se ve el nivel de conversión obtenido según el tramo de los puntajes ponderados obtenidos, notando que a mayores puntajes ponderados, el nivel de conversión de matrícula incrementa, variando de un 31% en el caso de los puntajes de los tramos más bajos (0 a 450 puntos ponderados), hasta sobre del 75% en los puntajes sobre los 600 puntos. Esta misma observación se da al disgregar por tipo de colegio.

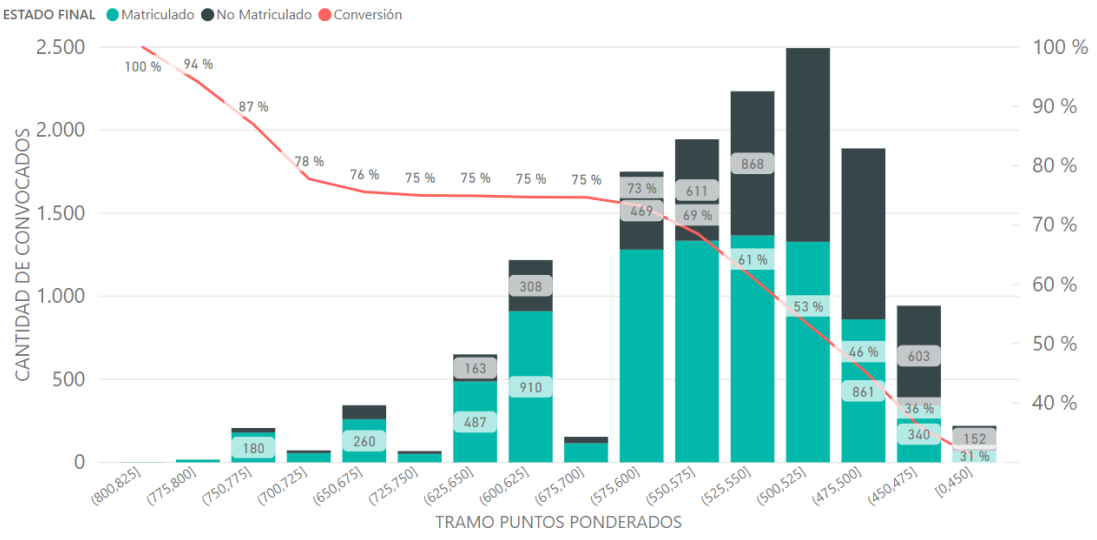


Imagen 20. Nivel de conversión de matrícula según tramo de puntaje ponderado.

Dentro de las acciones que utiliza la universidad como punto de contacto con sus alumnos, a traves de difusion, contacto telefonico o si los postulantes simulan beneficios, se puede observar que aquellos posibles postulantes que se ven enfrentados a esta interacción con la casa de estudios, muestran un mayor interes a matricularse en comparación a aquellos que no la tienen (Imagen 20,21,22), teniendo en algunos casos más de 10% de diferencia en los ratios de conversión entre los que si obtienen el contacto (valor = 1), contra los que no.

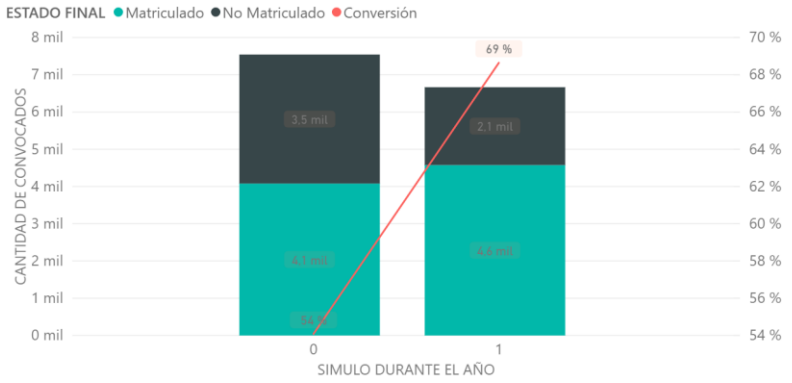


Imagen 21. Cantidad de matriculados que simularon durante el año..

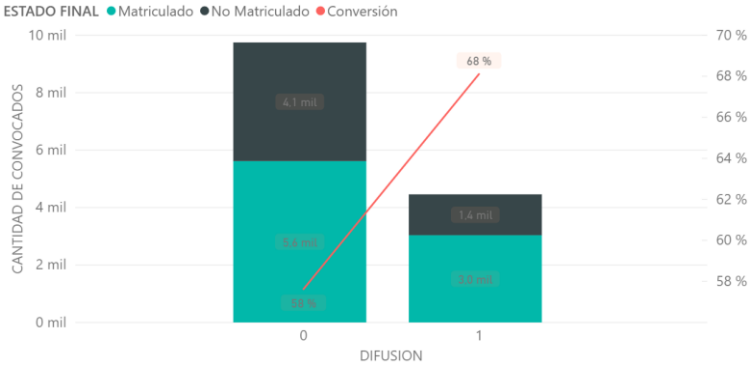


Imagen 22. Cantidad de matriculados participantes de difusión.

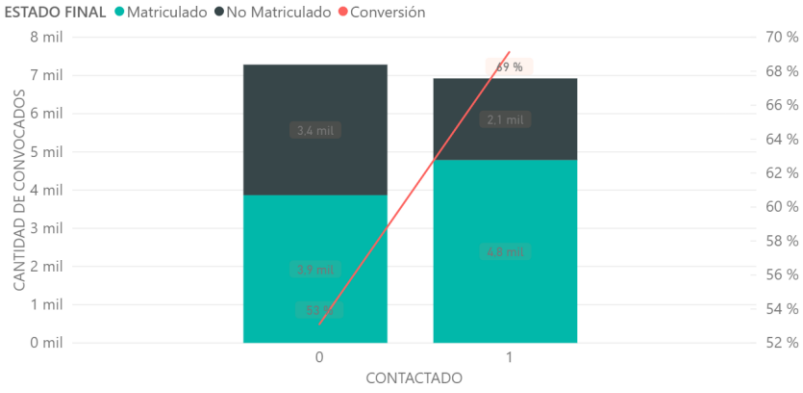


Imagen 23.Cantidad de matriculados contactados durante el año.

6.2.3. Convocados según tipo de egreso

Dentro del actual sistema de admisión, un postulante a la universidad puede rendir la PSU y guardar sus puntajes para postular al año siguiente, o volver a dar la prueba y postular con los resultados del año en el que obtuvo mejores puntajes. Es por lo anterior que se realiza una diferenciación en los tipos de convocados según su año de egreso y postulación en un proceso anterior, destacando la existencia de 3 grupos de convocados:

Grupo 1: Recién egresados del colegio que postulan.

Grupo 2: Egresados de otros años, que fue convocado en el proceso anterior.

Es así que se pueden tener personas que postulen un año, no se matriculen y al año siguiente realicen nuevamente el proceso de postulación.

Grupo 3: Egresado de otro año, no convocado dentro del proceso anterior.

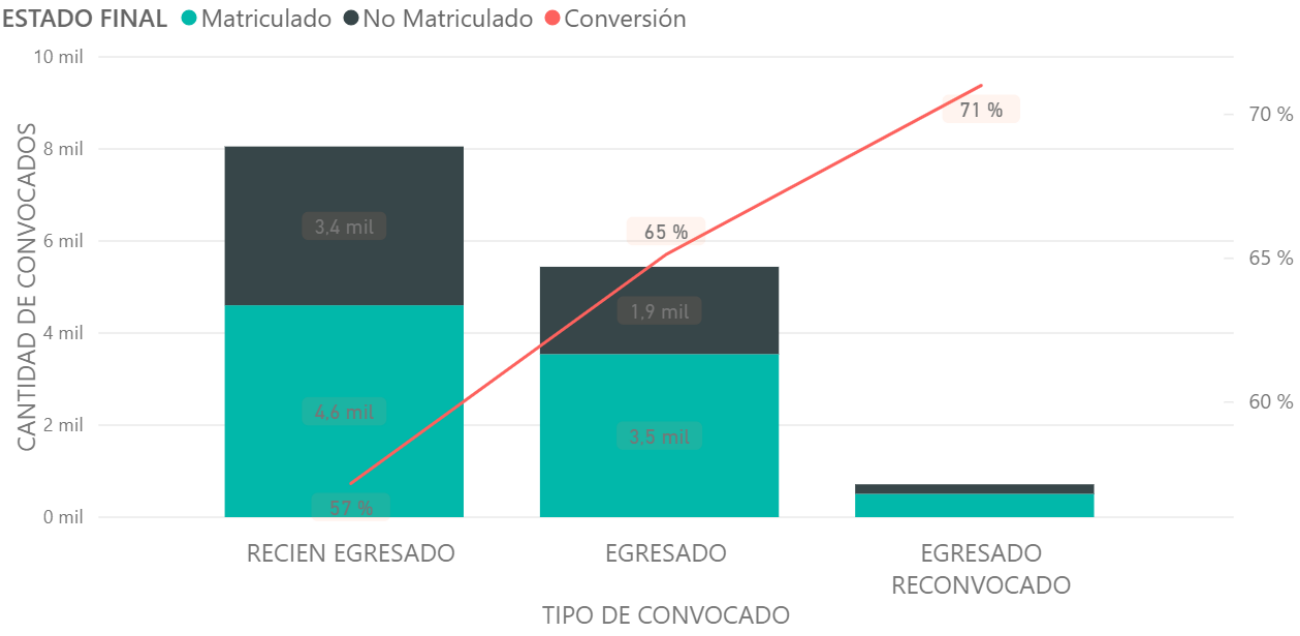


Imagen 24. Convocados diferenciados por grupo.

Dentro de la imagen 24, se puede notar que el grupo de los egresados reconvocados tiene un nivel de conversión de matrícula del 71% (el más alto dentro de todos los grupos) en contraste a un recién egresado, donde la conversión es del 57%. Tanto en los procesos 2018 y 2019 se puede notar que los egresados de años distintos al proceso de admisión entrante tienen mejores tasas de conversión que alumnos que están por primera vez en el proceso (Tabla 8,9,10).

Tabla 7. Conversión convocados en su mismo año de egreso.

Año	Convocados	Matriculados	Conversión
2018	7.320	3.658	50%
2019	8.053	4.604	57%

Tabla 8. Conversión convocado egresado en otro año, no convocado anteriormente.

Año	Convocados	Matriculados	Conversión
2018	5.660	3.254	57%
2019	5.440	3.543	65%

Tabla 9. Conversión convocado egresado en otro año, convocado anteriormente.

Año	Convocados	Matriculados	Conversión
2018	575	395	69%
2019	714	507	71%

6.3 Convocados no matriculados

Estudiar a aquellos convocados no matriculados se considera fundamental a la hora del estudio del fenómeno, ya que el entender las causas no previsibles de los modelos es una fuente de información extra totalmente aleatoria y ya fuera de los alcances y esfuerzos realizados dentro de este trabajo.

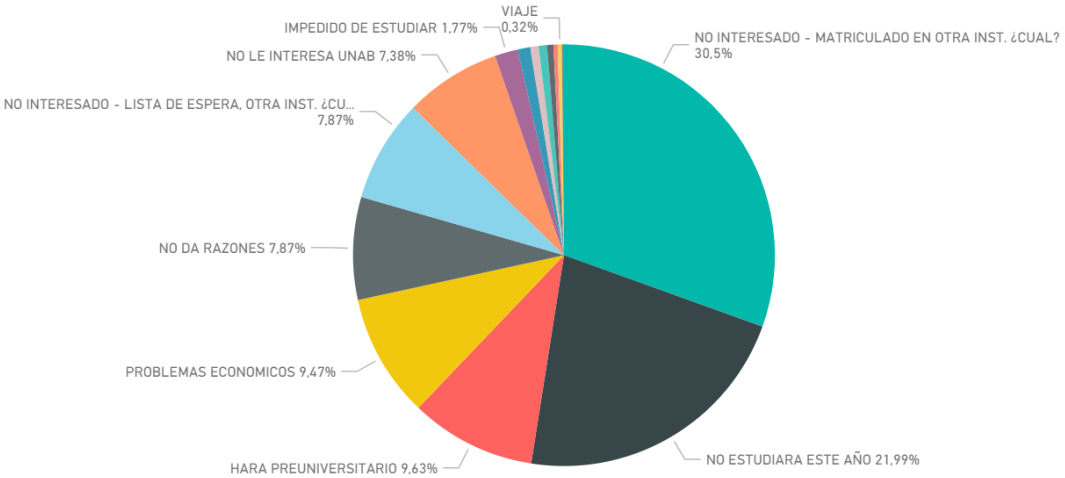


Imagen 25. Razones de no matrícula, proceso 2019.

Dentro de las causas más comunes de no matricularse (Imagen 18), se destaca como principal causa (38%) de no haberse matriculado, el haberse inscrito dentro de otras instituciones (matriculado en otra institución o en lista de espera), seguida por no estudiarán este año o harán preuniversitario (31%) y problemas económicos (10%). Esta información es importante de conocer pues aporta una idea de las problemáticas fuera de los alcances del proceso de matrículas que terminan afectando la decisión final de los convocados seleccionados.

Un análisis con mayor tamaño de muestra e información permitiría conocer si existen carreras que puedan ser homologables, ya que como expone Patricio Meller, “los alumnos más que escoger una universidad, escogen una facultad en la cual estudiar”, lo que da a pensar que existan carreras vistas como homologables entre estudiarlas en una institución u otra, basado en esta elección de carrera-universidad.

7. MODELOS DE DATA MINING

Dentro de esta sección se describen las variables utilizadas dentro de las etapas del proceso de minería de datos, siguiendo la metodología del KDD presentada en la sección 4.3.

Para la construcción de los modelos se hace uso de R para la calibración y estudio de los resultados de los distintos modelos de clasificación (Regresión logistca, Random Forest, SVM y Naive Bayes).

Para la fase de calibración y testeo, se separa la base de datos utilizado la información del proceso 2018 como set de entrenamiento y los datos del proceso 2019 como set de testeo. Cada uno de las bases de datos tiene cerca de 20.000 observaciones de los alumnos convocados en los procesos de admisión de cada año.

7.1. Definición del problema a tratar

El caso de interés tiene relación a predecir si el convocado se matriculará o no a partir de diversas variables obtenidas a lo largo del proceso de admisión. Dentro de la bibliografía consultada, se considera adecuado el uso de modelos de clasificación en donde la variable de interés queda representada por:

$$Matriculado = \begin{cases} 1, & Alumno\ convocado\ concretó\ su\ matricula \\ 0, & \sim \end{cases}$$

La variable a predecir ("Matriculado") se considera para los primeros 3 días de matrículas del proceso del SUA, sin considerar la existencia del efecto de temporalidad entre los distintos días.

7.2 Variables utilizadas

Las variables escogidas para la modelación del problema de clasificación de convocados según si se matricularán o no, fueron escogidas tomando en consideración el indicador de conversión de matrícula contrastado con variables socioeconómicas, postulación según el tipo de beneficios obtenidos, contacto con la universidad, entre otras. Como principales filtros de información se considera si el seleccionado egresa el mismo año del proceso y si este postuló o no en el pasado a la universidad estudiada.

7.2.1 Variables de beneficios

Dado que los beneficios a los que opten los convocados son una parte importante a la hora de matricularse o no dentro de la universidad, es que se utiliza la información de obtención de distintas becas en forma de variables binarias.

$$BECA\ EXTERNA = \begin{cases} 1, & El\ postulante\ tiene\ becas\ del\ Estado \\ 0, & \sim \end{cases}$$

$$CAE = \begin{cases} 1, & El\ postulante\ puede\ optar\ al\ CAE \\ 0, & \sim \end{cases}$$

$$BECA\_UNIVERSIDAD = \begin{cases} 1, & El\ postulante\ es\ preseleccionado\ recibe\ beca\ interna \\ 0, & \sim \end{cases}$$

7.2.2 Variables de postulación

Tomando en cuenta características de postulación, se crean variables relacionadas a la postulación del alumno, en donde se toma como criterio que si el convocado opta a una carrera dentro de sus 3 primeras opciones de las 10 existentes, entonces existe un mayor interés del convocado por esta casa de estudios.

$$PRIMERA\_PREF = \begin{cases} 1, & Si\ el\ alumno\ elige\ la\ carrera\ entre\ sus\ primeras\ 3\ opciones \\ 0, & \sim \end{cases}$$

7.2.3 Variables contacto con la universidad

La información de las bases de datos de aquellas instancias en que la universidad tiene contacto con el alumno, como lo son el uso del simulador de beneficios, las actividades (charlas, ensayos) y los llamados de call-center llamadas durante el periodo de conocimiento de los puntajes PSU es transformada y utilizada como variables binarias.

$$CONTACTADO\_TEL = \begin{cases} 1, & El\ call-center\ logra\ contactar\ al\ alumno \\ 0, & \sim \end{cases}$$

$$DIFUSION = \begin{cases} 1, & El\ alumno\ tiene\ contacto\ con\ las\ actividades\ de\ la\ universidad \\ 0, & \sim \end{cases}$$

$$SIMULO\_ANTES = \begin{cases} 1, & Alumno\ que\ simuló\ durante\ el\ año \\ 0, & Alumno\ que\ no\ simuló\ durante\ el\ año \end{cases}$$

7.2.4 Variables socioeconómicas

Dentro de la información socioeconómica declarada por el convocado, se utiliza la ubicación de su colegio para calcular la distancia existente entre el colegio de éste y la sede a la cual postula. La variable **DISTANCIA** es normalizada y transformada logarítmicamente de modo que no pondere más que el resto al calibrar los modelos.

Si bien la variable decil captura la información socioeconómica de las personas en el proceso, esta puede estar sesgada "hacia abajo" en casos que los alumnos mientan en sus papeles de postulación para optar a mejores beneficios de becas, aun así, esta variable es utilizada para calcular la cuota estimada que debiese pagar. Para esto se transforma DECIL a su monto de dinero correspondiente[4] tomando el valor promedio de los tramos en los cuales se encuentre el valor, para finalmente crear respecto a las cuotas que debe pagar el convocado en cada uno de los semestres:

Cuota estimada = Arancel Real / Monto dinero correspondiente al decil

[4] Según los datos presentados en el anexo de tabla de conversión

Al tratarse de una variable continua, al igual que distancia, es normalizada y transformada logarítmicamente.

La variable DECIL, es categorizada tomando 3 niveles (decil 1 al 3, decil 4 al 6, decil 7 al 10), donde cada uno de los tramos corresponde a una variable binaria.

Los puntajes ponderados por los convocados son categorizados en distintos tramos, creando 5 variables que indican si el alumno está dentro o no de esos puntajes, ya que como se vio en el análisis exploratorio, a mayor puntaje, mayor nivel de conversión de matrícula.

$TIPO\ DE\ TRAMO$: Indica 5 tramos de puntajes ponderados: (0-500 puntos), (500-550 puntos), (550-600 puntos), (600-700 puntos), (700-850 puntos).

La variable dependencia del colegio, se recodifica en 3 variables binarias:

$$PARTICULAR = \begin{cases} 1, si\ DEPENDENCIA\ =\ 1 \\ 0, \quad \sim \end{cases}$$

$$P\_SUB = \begin{cases} 1, si\ DEPENDENCIA\ =\ 2 \\ 0, \quad \sim \end{cases}$$

$$MUNICIPAL^{5} = \begin{cases} 1, si\ DEPENDENCIA\ =\ 3 \\ 0, \quad \sim \end{cases}$$

7.2.5 Indicadores de rendimiento del negocio

Como forma de incluir el comportamiento histórico tanto de la conversión de los colegios, como de las carreras, se crean las siguientes variables:

$CONVESION\_COLEGIO$ = Nivel de conversión promedio que ha tenido el colegio del convocado en los 2 procesos anteriores, para alumnos en el primer periodo de selección.

$CONVESION\_CARRERA$ = Nivel de conversión promedio que ha tenido la carrera dentro de los 2 procesos anterior. Notar que para carreras nuevas se toma solamente la información de un proceso de postulación.

[5] Esta variable es tomada dentro de la línea base de su categoría, por lo cual no se incluye dentro de los modelos.

7.3 Modelos

Dentro de la fase de construcción de los modelos, se utilizan los datos del proceso de admisión 2018 como datos de entrenamiento, en donde se utiliza un set de 19 variables para calibrar los modelos (15 variables binarias y 4 variables continuas). Dentro del set de entrenamiento no se considera a los convocados dentro "carreras de alta demanda" (por no ser de interés de estudio en las predicciones, además de ser datos *outliers* al compararlos con el resto de convocados (por sus altos puntajes principalmente)).

Dentro de este trabajo se siguieron 2 formulaciones, una en que se utilizan directamente todos los convocados para el entrenamiento de los modelos y otra en que estos son diferenciados y el resultado de sus modelos son agregados.

Al diferenciar los modelos utilizados según el tipo de respuesta que entregan:

- Modelo con respuesta continua: Dentro de estos modelos está el de regresión logística, el cual entrega una respuesta continua entre 0 y 1. Para calcular el punto donde el modelo entrega la mejor clasificación, se calcula "el punto de corte" optimo donde se alcanzan los mejores resultados de la clasificación, utilizando para esto el set de entrenamiento.
- Modelo con respuesta discreta: Dentro de los modelos utilizados para la clasificación se puede encontrar Random Forest, Super vector Machine y Naive Bayes. Estos modelos entregan directamente una respuesta discreta (0 o 1).Por cómo están formulados, se puede tratar de mejorar sus resultados al variar sus hiperparámetros (RF y SVM).

7.3.1 Modelos iniciales

En principio son los modelos son calibrados utilizando dentro del set de entrenamiento con los 3 tipos de convocados sin diferenciarlos. Para ser justo en la forma de comparar los modelos, se utiliza el mismo set de variables dentro de cada uno de los modelos calibrados.

7.3.2 Modelos agregados

Según lo expuesto en el punto 6.2.3(Convocados según tipo de egreso), se denota un comportamiento distinto entre aquellos convocados egresados el mismo año en que rinden el proceso y aquellos que salieron años antes (y ya hayan o no sido convocados a la universidad), es por esto que se sigue esta fase de modelamiento en que se calibran cada uno de los modelos para cada uno de los 3 tipos de convocados y sus resultados son agregados para compararlo con los modelos iniciales.

La división del total de datos de entrenamiento entre los distintos tipos de convocados queda cercano a una proporción 7:1:4 para los sets de convocados recién egresados, egresados otros años (convocados antes) y egresados otros años (no convocados antes) respectivamente.

7.4 Calibración e implementación del modelo

Se utiliza el software estadístico R para la construcción de los distintos modelos. Utilizando los modelos supervisados de machine learning: logistic regresión, random forest, SVM y Naive Bayes.

8. RESULTADOS

Dentro de este capítulo muestran los principales resultados del testeo de los distintos modelos calibrados. La elección del mejor modelo es realizada considerando la comparación entre los niveles de Accuracy, Specificity, Sensibility, %Positive predicted y %Negative predicted que alcancen los modelos dentro del set de testeo (proceso de admisión 2019).

Es importante destacar que, dentro de las características de los datos, ambas clases (entre matriculados y no matriculados) se encuentran en porcentajes cercanos a un 55% y 45% aproximadamente.

8.1 Métricas de los modelos

Dentro de los modelos, como primera medida de comparación, es utilizada la métrica de Accuracy como primer comparador, además de utilizar Sensibility y POS_PRED, ya que se quiere tener un grado alto de aciertos sobre los no matriculados (dentro de este caso, se considera clase positiva (0 : "no matriculado")).

Notar que los modelos agregados corresponden a la unión de los 3 modelos diferenciados por tipo de convocado.

Tabla 10. Métricas modelos agregados

	Accuracy	Sensitivity	Specificity	Pos Pred	Neg Pred
Logit	73,6%	60,9%	81,7%	67,9%	76,6%
Random Forest	73,0%	55,6%	84,0%	68,8%	74,8%
SVM	72,4%	64,8%	79,0%	72,8%	72,0%
Naive Bayes	72,6%	49,6%	87,2%	71,1%	73,1%
Mediana	72,8%	58,3%	82,8%	70,0%	74,0%

En general, se puede apreciar dentro de la comparación de los modelos agregados, que en niveles de Accuracy de los distintos modelos ronda al 73%, teniendo como ganadores dentro de esta métrica a los modelos Logit y Random Forest, los cuales alcanzan niveles de sobre un 73,6% y 73% respectivamente (Tabla 10), siguiendo con los mejores niveles de *Sensitibity* nuevamente por Logit y SVM. En cuanto a niveles de "*Pos Pred*" los resultados de los 4 modelos son similares, sin existir un ganador que resalte.

Al revisar los resultados de los modelos disgregados (Tabla 11), se puede notar que los modelos representados como A, B y C corresponden a recién egresados, egresados otros años (convocados antes) y egresados otros años (no convocados antes). En general los niveles de Sensitivity decaen para los grupos B (el grupo de menor tamaño en los datos) al compararlo con lo que sucede con los otros modelos, mostrando que el realizar un desglose dentro del modelamiento de este grupo, serviría para identificar mucho mejor a los convocados que si se matricularan.

Tabla 11. Métricas modelos disgregados y agregado por tipo de convocado.

	Accuracy	Sensitivity	Specificity	Pos Pred	Neg Pred
Logit A	72,3%	58,1%	82,9%	71,6%	72,7%
Logit B	76,7%	39,8%	91,7%	66,1%	78,9%
Logit C	76,0%	55,5%	86,8%	69,0%	78,6%
Logit Agregado	**73,9%**	**56,5%**	**85,0%**	**70,6%**	**75,4%**
RF A	71,5%	58,2%	81,3%	69,9%	72,4%
RF B	75,0%	32,5%	92,3%	63,2%	77,0%
RF C	74,6%	48,1%	88,7%	69,3%	76,3%
RF Agregado	**72,9%**	**53,8%**	**85,0%**	**69,5%**	**74,3%**
SVM A	70,4%	52,8%	83,5%	70,5%	70,4%
SVM B	77,4%	38,3%	93,3%	69,9%	78,8%
SVM C	75,8%	49,8%	89,6%	71,7%	77,1%
SVM Agregado	**72,8%**	**51,2%**	**86,6%**	**70,8%**	**73,6%**
NB A	70,5%	50,9%	85,0%	71,6%	70,0%
NB B	75,0%	41,3%	88,7%	59,9%	78,7%
NB C	75,1%	49,7%	88,5%	69,6%	76,9%
NB Agregado	**72,4%**	**50,2%**	**86,6%**	**70,5%**	**73,2%**

Tabla 12. Comparación métricas modelo logit

Métrica	LOGIT	LOGIT AGREGADO
ACCURACY	73,6%	73,9%
SENSITIVITY	60,9%	56,5%
SPECIFICITY	81,7%	85,0%
POS PRED	67,9%	70,6%
NEG PRED	76,6%	75,4%

Tabla 13. Comparación métricas modelo RF.

Métrica	RF	RF AGREGADO
ACCURACY	73,0%	72,9%
SENSITIVITY	55,6%	53,8%
SPECIFICITY	84,0%	85,0%
POS PRED	68,8%	69,5%
NEG PRED	74,8%	74,3%

Tabla 14. Comparación métricas modelo SVM

Métrica	SVM	SVM AGREGADO
ACCURACY	72,4%	72,8%
SENSITIVITY	64,8%	51,2%
SPECIFICITY	79,0%	86,6%
POS PRED	72,8%	70,8%
NEG PRED	72,0%	73,6%

Tabla 15. Comparación métricas modelo Naive Bayes

Métrica	NB	NB AGREGADO
ACCURACY	72,6%	72,4%
SENSITIVITY	49,6%	50,2%
SPECIFICITY	87,2%	86,6%
POS PRED	71,1%	70,5%
NEG PRED	73,1%	73,2%

En general, se puede apreciar que dentro de los indicadores, el logit y SVM Agregados muestra mejoras en sus métrica de Accuracy con deterioro en la Sensitivity, mostrando que en estos casos estos modelos sirven mejor al entrenarlos sin hacer diferenciación entre los convocados. La inspección inicial, permite apreciar que los modelos con mejores resultados dentro de la discriminación de clases son Logit y SVM.

Uno de los defectos de la metodología usada para la disgregación de los modelos está en utilizar el mismo set de variables para calibrar a los modelos. Es posible que de disminuir el número de estas tenga algún efecto en mejorar los indicadores de alguno de los modelos disgregados, ya que características como el contactarse con el convocado pareciera ser menos importante en aquellos casos de ya egresados del colegio en comparación a los recién egresados en el año actual del proceso.

8.2 Evaluación económica del uso de los modelos

Con el fin de evaluar cuál de los modelos da mejores resultados a la hora de predecir que convocado se matriculará o no durante el proceso de monitoreo de matrículas se propone evaluar la performance de estos tomando los resultados de una campaña simulada en la que se ofrecen distintos porcentajes de beca (con distintos porcentajes de aceptación de esta). La justificación de la elección del mejor modelo dentro de las campañas simuladas se realiza contrastando sus resultados económicos contra un modelo perfecto y otro aleatorio, bajo los supuestos de llamar al 30% de aquellos alumnos clasificados como no matriculados por los distintos modelos.

Modelo perfecto:

Dentro de un escenario de modelo ideal, que discrimina perfectamente a los que se matriculan y no se matriculan, siendo un oráculo perfecto.

Modelo Aleatorio:

Dentro de una campaña, predecir lanzando una moneda al aire si el convocado se matriculará o no (predice con un 50% de probabilidad en ambos resultados).

La evaluación económica del uso de los modelos presentados toma como base de comparación los resultados reales del proceso de matrículas en el proceso de admisión 2019.

Las siguientes métricas son usadas para encontrar la efectividad de las distintas acciones recomendadas.

- Costo otorgado en becas: Son utilizados calculando el monto de beca ofrecido a aquellos convocados que acepten y terminen matriculándose.
- Costo otorgado extra en becas: Calculado como la cantidad extra de beca que se le daría a un clasificado como no matriculado, pero que su resultado real es matriculado.
- Neto campaña modelo: Calculado como la diferencia entre el beneficio simulado por concepto de arancel, menos el costo otorgado en becas indicadas por el modelo.

Para la evaluación económica, se toman como supuestos que se llama a una muestra al azar de 30% de la predicción del modelo catalogados como no matriculados, con un ofrecimiento de una beca de un 25% de beca de arancel[6], con recepciones positivas de un 30%, 20% y 1% del total de los convocados contactados a matricularse luego de haber sido contactado. Según lo expuesto dentro del punto 6.3 es razonable considerar que cerca del 20% de los convocados no matriculados podría matricularse. Al variar el porcentaje de respuestas positivas por parte de los clasificados como no matriculados se tienen los siguientes escenarios:

Escenario Pesimista

Este escenario considera una respuesta positiva a matricularse de un 1% de los contactados.

Se nota que se puede obtener entre 15 y 27 convocados más.

Escenario Medio

Este escenario se considera más cercano a la realidad, en donde existe una tasa de aceptación del 20% de los contactados, otorgando beneficios superiores a los $500.000.000 solo por concepto de aranceles. Si bien de las becas seleccionadas, reparte cerca del 30% como costo de oportunidad, podrían tenerse cerca de **190 convocad**os[7] que pasen a ser matriculados.

Escenario optimista

Dentro del escenario que tuviese mejor recepción al tener un 35% de aceptación por parte de los contactados, en donde se obtienen beneficios superiores a los $980.000.000, matriculando entre **318 y 372 convocados**.

[6] Máxima beca ofrecida por política de la universidad dentro del 2° proceso de matrículas.

[7] Esto es cercano a obtener un 1% más centro de conversión en matricula.

Se comprueba la factibilidad económica y operacional de la propuesta en los tres escenarios propuestos. En general se puede observar dentro de los escenarios, es que se obtienen beneficios de la campaña que son entre 2 y 3 veces superiores al modelo aleatorio y cerca de la mitad de lo que supondría tener un modelo perfecto.

El análisis de sensibilidad de los ratios entre las distintas recepciones de la campaña denota que es conveniente bajo los supuestos de una campaña usando alguno de los modelos descritos como herramienta al apoyo de las campañas realizadas durante el monitoreo de las matrículas, lo que permitirían aumentar la conversión de matrícula, dada las actuales tendencias de los convocados. (Ver sección 6. Convocados no matriculados). Las tablas 12,13 y 14 muestran los ingresos y costes de los experimentos simulados. Es importante notar que en el proceso de monitoreo de matrículas, son llamadas cerca de 3100 personas por concepto de campaña de comunicación de beneficios, por lo cual el llamar a cerca del 30% de la muestra de convocados identificados con baja propensión a la matricula, está dentro de las actuales capacidades operacionales.

Notar que si bien los costos del call-center no están considerados dentro de la evaluación económica, ésta fluctúa en cerca de $5 millones en un escenario en que se llamase a la totalidad de los convocados (Despreciable con relación al ingreso percibido por aranceles de una carrera, en donde con 3 matriculados adicionales se costea la operación). La campaña considera call-center como la opción más cara, pero también están opciones como correos o SMS para la comunicación con los convocados.

Tabla 16. Resultados evaluación económica con tasa de respuesta del 35%.

35 % DE RESPUESTA	MODELO PERFECTO	MODELO AZAR	LOGIT MIX	LOGIT	RF	SVM	NB
	Promedio	Promedio	Promedio	Promedio	Promedio	Promedio	Promedio
Ingreso campaña vía modelo	$ 2.774.534.200	$ 1.397.345.400					
Monto otorgado en becas	$ 629.942.200	$ 820.701.700					
Monto becas extra	$ -	$ 503.248.100					
% Monto de becas extra	0%	61%					
Ingreso neto campaña	**$ 2.144.592.000**	**$ 576.643.700**					
Convocados a contactar	1664,6	2135					
Matriculas esperadas	587,2	294,6					

Tabla 17.Resultados evaluación económica con tasa de respuesta del 20%.

20 % DE RESPUESTA	MODELO PERFECTO	MODELO AZAR	LOGIT MIX	LOGIT	RF	SVM	NB
	Promedio	Promedio	Promedio	Promedio	Promedio	Promedio	Promedio
Ingreso campaña vía modelo	$ 1.701.273.800	$ 786.341.600					
Monto otorgado en becas	$ 386.361.400	$ 471.757.350					
Monto becas extra	$ -	$ 293.183.150					
% Monto de becas extra	0%	62%					
Ingreso neto campaña	**$ 1.314.912.400**	**$ 314.584.250**					
Convocados a contactar	1641	2113,8					
Matriculas esperadas	357	166,8					

Tabla 18.Resultados evaluación económica con tasa de respuesta del 1%.

1 % DE RESPUESTA	MODELO PERFECTO	MODELO AZAR	LOGIT MIX	LOGIT	RF	SVM	NB
	Promedio	Promedio	Promedio	Promedio	Promedio	Promedio	Promedio
Ingreso campaña vía modelo	$ 145.194.600	$ 73.570.800					
Monto otorgado en becas	$ 32.981.850	$ 44.762.650					
Monto becas extra	$ -	$ 28.066.600					
% Monto de becas extra	0%	63%					
Ingreso neto campaña	$ 112.212.750	$ 28.808.150					
Convocados a contactar	1638,8	2119,2					
Matriculas esperadas	30,8	15,8					

9. CONCLUSIONES

Se tomaron dos enfoques dentro de la creación de los modelos: entrenando los modelos con los datos agregados y otro en que se realizó la diferenciación por tipo de convocados agregando sus resultados para la comparación de los modelos. Tras comparar la performance de los modelos en el testeo de resultados y en la evaluación dentro de una campaña simulada, se puede notar que las métricas de evaluación son similares, posiblemente por el uso del mismo set de variables dadas en el entrenamiento. En los distintos enfoques adoptados el modelo logit resultó ganador.

A continuación, se presentan las siguientes conclusiones de este trabajo de título:

- Se cumplió el objetivo general del trabajo, entregando un modelo que permite identificar a los convocados con menor tendencia a matricularse, lo cual dentro de un caso de uso pesimista lograría una conversión adicional de un 1%, con beneficios esperados de $500 millones por concepto de aranceles.
- La comparación de escenarios simulados dentro de los cuales se ocuparían los modelos permite evaluar económicamente el valor del modelo como herramienta de gestión frente a casos en que se tomen decisiones con criterios arbitrarios, ya que aporta beneficios entre 2 y 3 veces mayores a un modelo aleatorio y cerca de la mitad de lo que haría un modelo con 100% de predicción.
- Es necesario hacer uso de varias métricas de evaluación de resultados al comparar los modelos, en donde para este caso, Accuracy necesita ser complementado por "Sensibility", pues dentro del objetivo del negocio, se quiere poder acertar más veces en los "casos positivos" de estudio. [8]
- El disgregar modelos acordes a distintos tipos de convocados puede provocar mejoras dentro de las métricas de evaluación de los modelos a nivel disgregado, sin embargo, el uso del mismo set de variables pusiese representar un elemento que afecte su poder predictivo.

[8] Dentro de esta memoria, el caso positivo es la clase = 0 (No matriculado).(Esto varía según el objetivo del proyecto).

10. BIBLIOGRAFIA

Libros, artículos y documentos

Barrientos, F. (2014). *Optimización de la oferta de becas en una institución de educación superior.* Santiago: Universidad de Chile.

Barrientos, F., & Rios, S. (2013). Aplicación de mineria de datos para predecir fuga de cliente en la industria de las telecomunicaciones. *Revista ingeniería de sistemas*, 73-107.

Chile, V. (18 de Octubre de 2016). *Mitos y Verdades sobre la Educación Superior Técnico Profesional.* Obtenido de https://www.youtube.com/watch?v=Ba1Q8q7B26g.

Espinace, B. (2014). *Desarrollo de un plan de difusión para una universidad privada perteneciente al sistema único de admisión.* Santiago: Universidad de Chile.

Meller, P. (2010). Carreras universitarias, rentabilidad, selectividad y discriminación. Santiago de Chile: Uqbar Editores.

Meller, P. (2011). *Universitarios, !el problema no es el lucro, es el mercado!* Santiago de Chile: Uqbar Editores.

Meza, L. (2016). *Estimación de ocupación de carreras universitarias para una universidad adscrita al sistema único de admisión.* Santiago: Universidad de Chile.

Ministerio de Desarrollo Social. (2018). *Informe de desarrollo social.* Obtenido de http://www.ministeriodesarrollosocial.gob.cl/storage/docs/Informe_de_Desarrollo_Social_2018.pdf

Núñez, C. (2013). *Modelo de competitividad para una universidad privada adscrita al sistema DEMRE.* Santiago: Universidad de Chile.

Nuñez, M. (2018). *Solución analítica para plan de difusión y aumento de postulaciones en universidad perteneciente al sistema único de admisión.* Santiago: Universidad de Chile.

Apéndice A

Estructura de datos

A.1. Estructura de datos consolidada.

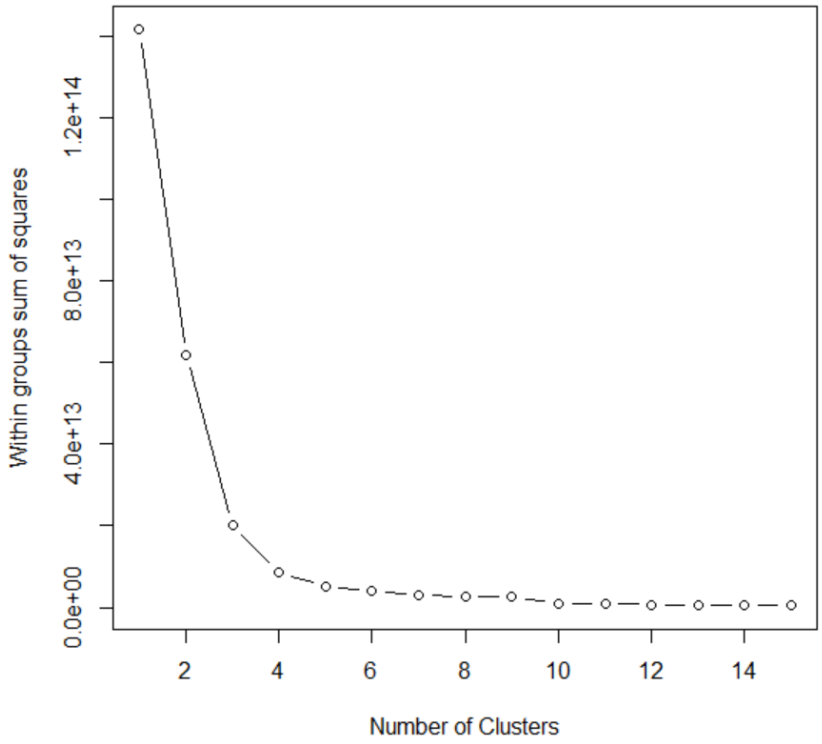
# Variable	Variable	Descripción	Valores
1	RUT (character)	Identificador único del estudiante sin digito verificador.	Ejemplo: “19867543”
2	ESTADO_POSTULACION (character)	Resultado de la postulación una vez conocidos los puntajes PSU.	Ejemplo: “CONVOCADO”,” LISTA ESPERA”.
3	AÑO_PROCESO (numeric)	Año del proceso DEMRE.	Ejemplo: 2018, 2019
4	SEXO (character)	Genero del postulante	Ejemplo: “Masculino”, ”Femenino”
5	AÑO_EGRESO_COL (numeric)	Año egreso del colegio del postulante	Ejemplo: 2018
6	COD_CARRERA (character)	Código de la carrera postulada	Ejemplo: “41067”
8	CAE (character)	Si el alumno está preseleccionado al CAE	Ejemplo: “CAE”,”NA”.
9	BECA_EXT (character)	Si el alumno está preseleccionado a otras becas del estado	Ejemplo: “JGM”, “BB”.
10	CAMPUS(character)	Campus elegido por el postulante	Ejemplo: ”REPUBLICA”.
11	FACULTAD (character)	Facultad de la carrera	Ejemplo: ”FACULTAD DE MEDICINA”
12	COMUNA_CAMPUS (character)	Comuna del campus	Ejemplo: ”SANTIAGO”
13	REGION_CAMPUS (numeric)	Código de la región del campus	Ejemplo: 13
14	CARRERA (character)	Nombre de la carera	Ejemplo: ”PSICOLOGIA”.
15	MIN_PTOS_PONDERADO (numeric)	Mínimo puntaje ponderado para postular	Ejemplo: 450
16	MIN_LENG_MATE (numeric)	Mínimo puntaje ponderado entre las psu de lenguaje y matemáticas	Ejemplo: 450
17	VACANTES_DEMRE (numeric)	Cantidad de vacantes disponibles para el primer proceso de matricula	Ejemplo: 75
18	COD_REG (numeric)	Código de la región del postulante	Ejemplo: 6
19	COD_COM (numeric)	Código de la comuna del postulante	Ejemplo: 1101
20	COMUNA (character)	Comuna del colegio del postulante	Ejemplo: ”RANCAGUA”
21	LAT_COL (numeric)	Latitud de la comuna del colegio del postulante	Ejemplo: -43,11

# Variable	Variable	Descripción	Valores
22	LON_COL (numeric)	Longitud de la comuna del colegio del postulante	Ejemplo: -33,03
23	LAT_U (numeric)	Latitud de la comuna del campus	Ejemplo: -43,11
24	LON_U (numeric)	Longitud de la comuna del campus	Ejemplo: -33,03
25	DECIL (numeric)	Decil del postulante	Ejemplo: 1,2,3,4,5,6,7,8,9,10
26	DEPENDENCIA (numeric)	Tipo de colegio del postulante	Ejemplo: 1: "Particular" 2: "Particular subvencionado" 3: "Municipal"
27	LENG_MAT	Puntaje ponderado entre las psu de lenguaje y matemáticas	Ejemplo: 555
28	PUNTAJE_POND	Puntaje ponderado obtenido	Ejemplo: 580
29	PREF_POSTULACION	Opción de elección de la carrera postulada	Ejemplo: 1,2,3,4,5,6,7,8,9,10
30	DISTANCIA	Kilómetros entre el colegio del postulante y la comuna a la que postula	Ejemplo: 300

Apéndice B

Resultados de Data-Mining

B.1 Elección de las cantidad de clases según algoritmo K-means (criterio del codo).



B.2 Centroides de los clusters entre carreras

CLUSTER	PTOS_PROMEDIO	MONTO ARANCEL	CONVERSION PRIOMEDIO	TOTAL CONVOCADOS
1	543,9	$ 4.848.229	0,53	195,1
2	530,9	$ 2.904.534	0,44	49,02
3	534,5	$ 3.428.050	0,43	56,04
4	677,02	$ 7.258.537	0,67	129,5
5	457,7	$ 4.070.064	0,55	94,6

B.3 Cantidad de elementos por cluster.

CLUSTER	TAMAÑO
1	30
2	36
3	24
4	6
5	46

B.4 Cluster de “carreras destacadas”.

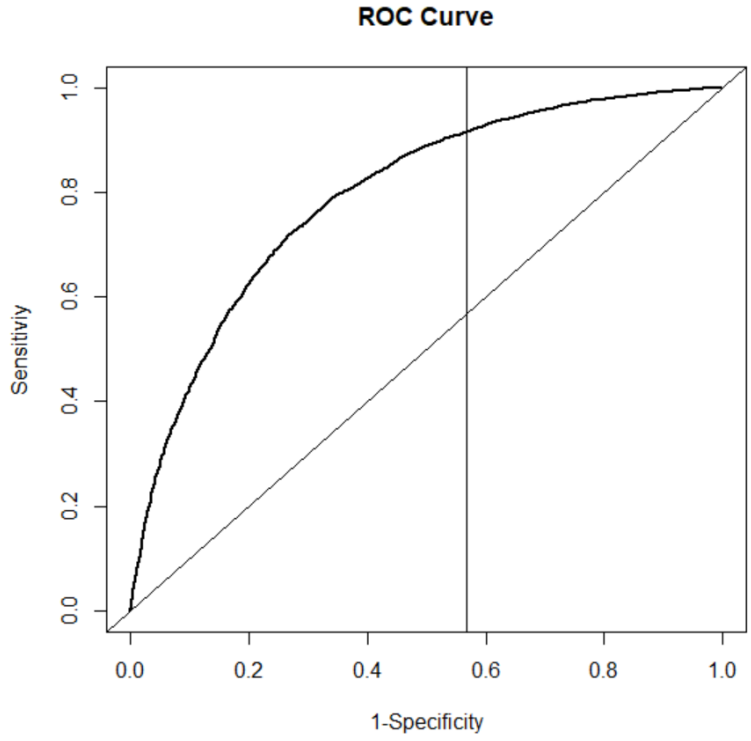
CODIGO	CARRERA	CAMPUS	PTOS PONDERADOS	DESV PTOS_POND	ARANCEL	CLUSTER
41056	MEDICINA	REPUBLICA	762,7	8,5	$ 7.990.000	4
41060	ODONTOLOGIA	REPUBLICA	595,0	45,1	$ 7.990.000	4
41095	MEDICINA	VINA DEL MAR	764,5	6,8	$ 6.790.000	4
41099	ODONTOLOGIA	VINA DEL MAR	599,5	55,4	$ 6.890.000	4
41127	ODONTOLOGIA	CONCEPCION	589,7	44,9	$ 6.890.000	4
41164	MEDICINA	CONCEPCION	750,7	7,5	$ 7.050.000	4

B.5 Modelo logit agregado

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.36713    0.18919 -12.512  < 2e-16 ***
BECA_EXTERNA1       0.20360    0.05539   3.676 0.000237 ***
CAE1                0.57690    0.05260  10.967  < 2e-16 ***
BECA_UNAB1          1.16035    0.05149  22.534  < 2e-16 ***
PRIM_PREF1          1.18231    0.06330  18.677  < 2e-16 ***
CONTACATADO_TELEFONO1  0.23171    0.05410   4.283 1.85e-05 ***
DIFUSION1           0.07916    0.04619   1.714 0.086541 .
SIMULO_ANTES1       0.10203    0.04998   2.042 0.041188 *
LOG_1_DISTANCIA    -0.25035    0.09632  -2.599 0.009347 **
CUOTA_INICIAL_LOG  -0.84278    0.22668  -3.718 0.000201 ***
TRAMO_DECILD4_6     0.21829    0.06685   3.266 0.001093 **
TRAMO_DECILD7_10    0.50318    0.11075   4.544 5.53e-06 ***
DEPENDENCIA2       -0.59573    0.06826  -8.728  < 2e-16 ***
DEPENDENCIA3       -0.83335    0.08093 -10.297  < 2e-16 ***
TIPO_TRAMO4(500,550]  0.31441    0.05824   5.398 6.73e-08 ***
TIPO_TRAMO4(550,600]  0.61305    0.06607   9.278  < 2e-16 ***
TIPO_TRAMO4(600,700]  0.71098    0.08014   8.872  < 2e-16 ***
TIPO_TRAMO4(700,850]  0.64431    0.32800   1.964 0.049488 *
CONV_RBD            0.28985    0.12221   2.372 0.017706 *
CONVERSION          1.79434    0.16427  10.923  < 2e-16 ***
MISMO_ANO_EGRESO1  -0.71288    0.04643 -15.355  < 2e-16 ***
POSTULA_ANTES1      0.24550    0.11533   2.129 0.033274 *
---
Signif. codes:  0 æ***Æ 0.001 æ**Æ 0.01 æ*Æ 0.05 æ.Æ 0.1 æ Æ 1
```

Los parámetros de este modelo dan cuenta de los efectos marginales de cada una de las variables en incrementar o disminuir la probabilidad que el convocado finalice su matrícula.

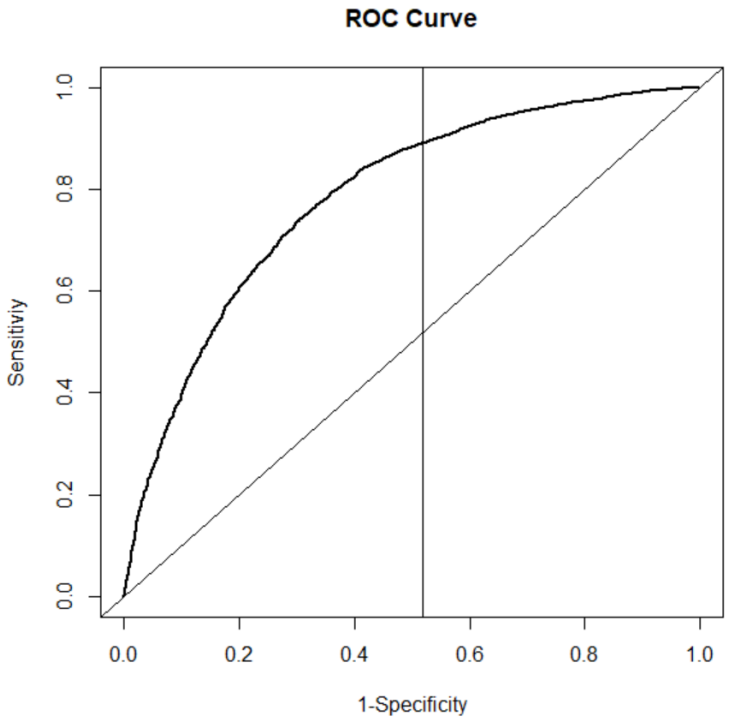
B.6 Cálculo del punto de corte set training(0,5668)



B.7 Modelo logit - modelo convocados recién egresados

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.86434    0.25517 -11.225  < 2e-16 ***
BECA_EXTERNA1         0.24515    0.07281   3.367 0.000759 ***
CAE1                  0.51436    0.07658   6.716 1.86e-11 ***
BECA_UNAB1            1.21573    0.06975  17.430  < 2e-16 ***
PRIM_PREF1            1.21156    0.09037  13.407  < 2e-16 ***
CONTACATADO_TELEFONO1 0.19599    0.06971   2.812 0.004931 **
DIFUSION1            -0.03574    0.06282  -0.569 0.569413
SIMULO_ANTES1         0.14351    0.06540   2.195 0.028195 *
LOG_1_DISTANCIA      -0.28968    0.13086  -2.214 0.026848 *
CUOTA_INICIAL_LOG    -1.03689    0.30278  -3.425 0.000616 ***
TRAMO_DECILD4_6       0.22035    0.09202   2.395 0.016635 *
TRAMO_DECILD7_10      0.42771    0.14612   2.927 0.003421 **
DEPENDENCIA2         -0.55525    0.09309  -5.965 2.45e-09 ***
DEPENDENCIA3         -0.80872    0.11219  -7.208 5.67e-13 ***
TIPO_TRAMO4(500,550]  0.20433    0.08062   2.535 0.011260 *
TIPO_TRAMO4(550,600]  0.54147    0.08923   6.069 1.29e-09 ***
TIPO_TRAMO4(600,700]  0.48182    0.10823   4.452 8.52e-06 ***
TIPO_TRAMO4(700,850]  0.13267    0.49012   0.271 0.786625
CONV_RBD              0.25264    0.16168   1.563 0.118152
CONVERSION            1.76563    0.21957   8.041 8.90e-16 ***
---
Signif. codes:  0 æ***Æ 0.001 æ**Æ 0.01 æ*Æ 0.05 æ.Æ 0.1 æ Æ 1
```

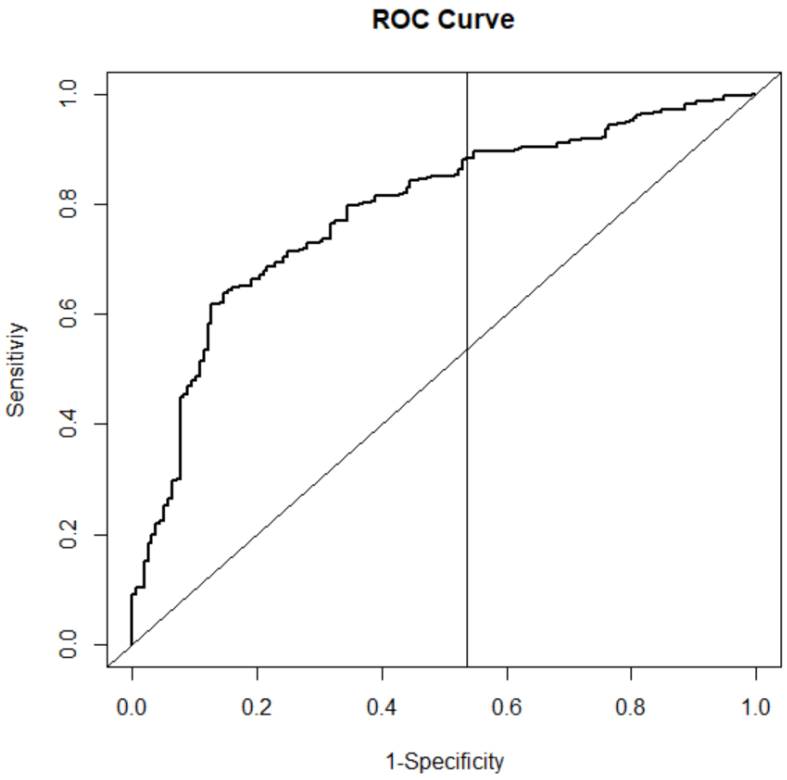
B.8 Calculo del punto de corte(0,5180)



B.9 Summary egresados años pasados, convocados anteriormente

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.70558    1.03003  -3.598 0.000321 ***
BECA_EXTERNA1        -0.18287    0.31300  -0.584 0.559045
CAE1                  0.28852    0.26718   1.080 0.280194
BECA_UNAB1            1.35097    0.27876   4.846 1.26e-06 ***
PRIM_PREF1            0.93615    0.31983   2.927 0.003422 **
CONTACATADO_TELEFONO1 0.29338    0.33460   0.877 0.380590
DIFUSION1             0.04614    0.23416   0.197 0.843803
SIMULO_ANTES1         0.27006    0.26863   1.005 0.314741
LOG_1_DISTANCIA      -0.92599    0.46856  -1.976 0.048128 *
CUOTA_INICIAL_LOG     2.05715    1.19084   1.727 0.084083 .
TRAMO_DECILD4_6       0.72882    0.38601   1.888 0.059011 .
TRAMO_DECILD7_10      1.53837    0.59908   2.568 0.010232 *
DEPENDENCIA2         -0.51102    0.33112  -1.543 0.122753
DEPENDENCIA3         -0.63904    0.40848  -1.564 0.117710
TIPO_TRAMO4(500,550] -0.04514    0.39573  -0.114 0.909195
TIPO_TRAMO4(550,600]  0.91106    0.41839   2.178 0.029440 *
TIPO_TRAMO4(600,700]  0.94705    0.48102   1.969 0.048973 *
TIPO_TRAMO4(700,850]  0.98189    1.46038   0.672 0.501359
CONV_RBD              0.13004    0.68883   0.189 0.850262
CONVERSION            2.29782    0.82608   2.782 0.005409 **
---
Signif. codes:  0 æ***Æ 0.001 æ**Æ 0.01 æ*Æ 0.05 æ.Æ 0.1 æ Æ 1
```

B.10 Importancia de las variables(0,5353)



B.11 Summary egresados años pasados, no convocados anteriormente

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.56770    0.29910  -8.585  < 2e-16 ***
BECA_EXTERNA1        0.20638    0.09040   2.283 0.022428 *
CAE1                 0.66209    0.07667   8.636  < 2e-16 ***
BECA_UNAB1           1.07936    0.08123  13.288  < 2e-16 ***
PRIM_PREF1           1.17617    0.09418  12.489  < 2e-16 ***
CONTACATADO_TELEFONO1  0.28796    0.09058   3.179 0.001478 **
DIFUSION1            0.22219    0.07241   3.069 0.002150 **
SIMULO_ANTES1        0.02987    0.08222   0.363 0.716370
LOG_1_DISTANCIA     -0.12517    0.15145  -0.826 0.408545
CUOTA_INICIAL_LOG   -0.77306    0.36550  -2.115 0.034423 *
TRAMO_DECILD4_6      0.20903    0.10287   2.032 0.042145 *
TRAMO_DECILD7_10     0.60463    0.18068   3.346 0.000818 ***
DEPENDENCIA2        -0.68692    0.10817  -6.350 2.15e-10 ***
DEPENDENCIA3        -0.90707    0.12436  -7.294 3.01e-13 ***
TIPO_TRAMO4(500,550]  0.47799    0.08717   5.484 4.17e-08 ***
TIPO_TRAMO4(550,600]  0.65177    0.10414   6.259 3.88e-10 ***
TIPO_TRAMO4(600,700]  1.01750    0.12819   7.938 2.06e-15 ***
TIPO_TRAMO4(700,850]  1.12164    0.48836   2.297 0.021634 *
CONV_RBD             0.41205    0.19683   2.093 0.036307 *
CONVERSION           1.78182    0.26290   6.778 1.22e-11 ***
---
Signif. codes:  0 æ***Æ 0.001 æ**Æ 0.01 æ*Æ 0.05 æ.Æ 0.1 æ Æ 1
```

B.12 Importancia de las variables (0,5554)

B.13 Pruebas de train y testeo

(a) Métricas en set de entrenamiento(Modelos diferenciados)(1:Convocados recién egresados;2= Convocados anteriormente, 3= nunca convocado)

	ACCURACY	SENSITIBITY	SPECIFICITY	POS PRED	NEG PRED	
Set de Train	72,3%	58,1%	82,9%	71,6%	72,7%	Modelo LR (1)
	76,7%	39,8%	91,7%	39,8%	91,7%	Modelo LR(2)
	76,0%	55,5%	86,8%	55,5%	86,8%	Modelo LR(3)
	71,4%	58,3%	81,2%	69,7%	72,4%	Modelo RF (1)
	76,1%	34,5%	93,1%	34,5%	93,1%	Modelo RF(2)
	75,0%	50,4%	88,0%	50,4%	88,0%	Modelo RF(3)
	70,4%	62,2%	78,6%	74,7%	67,2%	Modelo SVM (1)
	75,5%	52,2%	86,3%	52,2%	86,3%	Modelo SVM(2)
	73,3%	63,2%	81,0%	63,2%	81,0%	Modelo SVM(3)
	70,7%	68,9%	72,6%	71,8%	69,6%	Modelo NB (1)
	74,6%	57,3%	82,7%	57,3%	82,7%	Modelo NB(2)
	73,3%	68,0%	77,3%	68,0%	77,3%	Modelo NB(3)

(a) Métricas en set de testeo (Modelos diferenciados)(1:Convocados recién egresados;2= Convocados anteriormente, 3= nunca convocado)

	ACCURACY	SENSITIBITY	SPECIFICITY	POS PRED	NEG PRED	
Set de test	71,8%	70,0%	73,7%	73,0%	70,7%	Modelo LR (1)
	76,9%	57,3%	86,0%	57,3%	86,0%	Modelo LR(2)
	74,6%	70,8%	77,5%	70,8%	77,5%	Modelo LR(3)
	71,4%	58,3%	81,2%	69,7%	72,4%	Modelo RF (1)
	76,1%	34,5%	93,1%	34,5%	93,1%	Modelo RF(2)
	75,0%	50,4%	88,0%	50,4%	88,0%	Modelo RF(3)
	70,4%	52,8%	83,5%	70,5%	70,4%	Modelo SVM (1)
	77,4%	38,3%	93,3%	38,3%	93,3%	Modelo SVM(2)
	75,8%	49,8%	89,6%	49,8%	89,6%	Modelo SVM(3)
	70,5%	50,9%	85,0%	71,6%	70,0%	Modelo NB (1)
	75,0%	41,3%	88,7%	41,3%	88,7%	Modelo NB(2)
	75,1%	49,7%	88,5%	49,7%	88,5%	Modelo NB(3)

Apéndice C

Códigos

C.1 Código calibración modelos data-mining de clasificación

```
rm(list=ls())

#carga de los datos

#############################################################
#
#     MODELO LOGIT
#
#############################################################

  set.seed(1011)
  # Fitting Logistic Regression to the Training set
  LR = glm(formula = MATRICULADO ~ .,
           family = binomial(link='logit'),
           data = training_set)

p1<-predict(LR,test_set[-1],type = "response")
  p2<-ifelse(p1>opt_t,1,0)
  p2<-as.factor(p2)

  #test.predicted.m1 <- predict(classifier, newdata = test_set[-1], type = "response")
  confusionMatrix(p2,as.factor(test_set$MATRICULADO))
  library(ROCR)

  #### MOSTRAR EL VALOR DEL AUC
  # model 1 AUC
  prediction(as.numeric(p2),as.numeric(test_set$MATRICULADO)) %>%
    performance(measure = "auc") %>%
    .@y.values

#############################################################
#
#     MODELO RANDOM_FOREST
#
#############################################################
{
  # TODAS LAS CARRERAS
  RF = randomForest(
    formula = MATRICULADO ~ .,
    data    = training_set,
    ntree= 100,split="gini",mtry=3,nodesize=5
  )
###USNDO EL TEST DATA
  p2<-predict(RF,test_set)
  confusionMatrix(p2, test_set$MATRICULADO)
  #
  #### MOSTRAR EL VALOR DEL AUC
  prediction(as.numeric(p2),as.numeric(test_set$MATRICULADO)) %>%
    performance(measure = "auc") %>%
    .@y.values
```

```
#############################################################
#
#     MODELO SVM
#
#############################################################
{
  #install.packages('e1071')
  library(e1071)

  SVM = svm(formula = MATRICULADO ~ .,
            data = training_set,
            type = 'C-classification',
            kernel = 'linear')

###USNDO EL TEST DATA
      p2<-predict(SVM,test_set)
      confusionMatrix(p2, test_set$MATRICULADO)
      #
      ###ver como es que iva ajustando el arbol

      #### MOSTRAR EL VALOR DEL AUC
      # model 1 AUC
      prediction(as.numeric(p2),as.numeric(test_set$MATRICULADO)) %>%
        performance(measure = "auc") %>%
        .@y.values

#############################################################
#
#     MODELO NAIVE BAYES
#
#############################################################

  #install.packages('e1071')
  library(e1071)
  NB = naiveBayes(x = training_set[-1],
                  y = training_set$MATRICULADO)

p2<-predict(NB,test_set)
  confusionMatrix(p2, test_set$MATRICULADO)
  #
  #### MOSTRAR EL VALOR DEL AUC
  # model 1 AUC
  prediction(as.numeric(p2),as.numeric(test_set$MATRICULADO)) %>%
    performance(measure = "auc") %>%
    .@y.values
```