

# Comparative Analysis of Classification Techniques to Select Potential Female Applicants to Computer Related Careers in Northern Chile

Atsuko Galaz-Alday  
Departamento de Ingeniería y Tecnologías  
Universidad de Tarapacá  
Iquique, Chile  
atsuko.galaz@alumnos.uta.cl

Jorge Díaz-Ramírez  
Departamento de Ingeniería y Tecnologías  
Universidad de Tarapacá  
Iquique, Chile  
jdiazr@academicos.uta.cl

Ximena Badilla-Torrico  
Departamento de Ingeniería y Tecnologías  
Universidad de Tarapacá  
Iquique, Chile  
xbadilla@academicos.uta.cl

**Abstract**— Computer-related careers have maintained the stigma of being mostly masculine. Currently, such careers are demanded in labour market, but there are not enough professionals to meet demand and less than 20% of the students enrolled in technology-related careers are women, according to the Chilean Higher Education Information Services. The absence of information that characterizes the women who enroll in the Computer related area in Chile is the main motivation for this work.

This study presents a comparison of the results of classification techniques for the data set of female students who choose Computer Science in universities belonging to the Council of Rectors of Chilean Universities (CRUCH), in order to identify relevant variables to choose this careers. School location, academic performance, and mother's education were relevant. The results of two resampling schemes for imbalanced classes are similar, however Naïve Bayes with undersampling obtained slightly more balanced results with Prediction of 61%.

**Keywords**— Data mining, Classification model, KDD, Gender.

## I. INTRODUCCIÓN

Durante años, la participación de las mujeres en áreas como Ciencia, Tecnología, Ingeniería y Matemáticas (por sus siglas en inglés, STEM), ha sido una parte fundamental para el desarrollo de los países [1]. A pesar de los estudios de este comportamiento, el fenómeno parece persistir y se acentúa en carreras relacionadas con tecnología, donde Chile no es la excepción. Según datos de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) [2], la participación femenina en las matrículas de Ingeniería, Manufactura y Construcción alcanzan un 19%, mientras que el promedio de la OCDE es de un 25% en el año 2015.

Por otra parte un estudio de Microsoft Chile [3], evidencia que al analizar la distribución regional de la matrícula femenina en informática y computación, la zona norte alcanza un máximo de un 4%, mientras que la región metropolitana concentra el 49% de este tipo de matrículas. En esta misma línea, una investigación propone que para aumentar la participación

femenina en estos campos existen dos desafíos importantes, donde son el reclutamiento y la retención de las estudiantes [4], de ahí la importancia de este trabajo.

El estudio de la baja representación femenina en el área de informática y computación ha adquirido relevancia tanto para la economía como también el contexto social de los países. Estas carreras están siendo productivas y se prevé que es una de las carreras que más aumento de empleo tendrá hasta 2028 [5]. También fomentar los avances importantes y la creatividad en esta área requieren de pensamiento “fuera de la caja”, que puede ser fomentado por la diversidad de género [6]. Por último, la elección de una carrera es un proceso que involucra diversas variables, donde la edad promedio al enfrentar este escenario puede ser temprana [7], por ende, enfocar los estudios a áreas en particular, sería beneficioso para caracterizarlas y poder en un futuro guiar mejor a las estudiantes.

Por lo anterior, la presente investigación tiene por objetivo identificar variables que caractericen a las estudiantes que ingresaron a carreras relacionadas con computación e informática en una de las zonas que posee menor porcentaje de matrículas de mujeres en estas carreras, a través de la comparación del rendimiento de las técnicas de clasificación aplicados a un conjunto de datos que consideró la selección de variables relevantes como el desempeño escolar, educación de los padres y entorno familiar de las estudiantes. Para lograr esto, se utilizó la metodología KDD y se compararon los modelos mediante una tabla de métricas.

## II. MARCO TEÓRICO

### A. Predictores de elección de una carrera profesional

De acuerdo con Camarena, González y Velarde [7] afirman que la elección de una carrera es un proceso de alta complejidad, debido a que la mayoría de los individuos que afrontan esta situación lo hacen a temprana edad y deben enfrentarse a una amplia oferta de programas académicos, de los cuales no poseen la información suficiente sobre las áreas que lo componen ni su perfil profesional. A pesar de ello, existen diversos estudios que tratan de explicar y predecir la elección de una carrera. Uno de

ellos es el modelo clásico de Eccles, el cual propuso que las variables como los estereotipos y las experiencias en el aula, afectan las expectativas de éxito (autoeficacia), que a su vez afectan la selección de cursos y opciones de carrera [8]. En esta misma línea también existe una teoría de elección de carreras, denominada “teoría de la carrera cognitiva social”, donde enfatiza la interrelación entre variables individuales, ambientales y de comportamiento que pueden predecir el interés y la elección de una carrera [9], [10].

Gámez y Marrero [11] analizaron las metas y motivos que tienen los estudiantes de primer semestre de Biología, Derecho y Psicología en una Universidad de España. A través de un análisis factorial, la investigación aborda motivos de afinidad e interés por las relaciones, de logro y prestigio, motivación orientada al poder, superación de problemas afectivos y motivación extrínseca. Si se analizan las variables propuestas por las teorías [8]–[11], éstas se enfocan en el aspecto psicológico, por lo cual, es complejo de abordar un análisis individual con variables como la autoeficacia, confianza y amistades, si se explora la disponibilidad de los datos de los estudiantes que están ingresando a la universidad.

Por otra parte, también existe otra investigación que aborda este tema desde otra perspectiva. Bravo y Vergara [9] enfatizan que el rendimiento en la escuela refleja en parte qué tipos de carreras podrían elegir el alumno, de este modo, por ejemplo, no se espera que estudiantes con bajo rendimiento en matemáticas y ciencias, entren a programas relacionados con ingeniería.

Otro estudio analiza la influencia de la situación socioeconómica y laboral de los padres sobre la elección de un programa de educación superior por género. Concluyen que, sin importar esta última variable, estudiantes con padres con cargos gerenciales, tienen mayor probabilidad de elegir Ingenierías y Ciencias. De hecho, es más probable que mujeres prefieran programas de Humanidades, Ciencias sociales o Ingeniería sobre el programa de Negocios, cuando el nivel socioeconómico no es muy alto [12].

Porter y Umbach [13] a través de un modelo Logit Multinomial, explican la elección de carrera por medio de variables demográficas, influencia de padres, resultados en la prueba de matemáticas, actitud verbal, entre otras. También existen otras variables como el ingreso familiar y el tipo de financiamiento de la educación superior [9], [10] y el nivel educativo de los padres [14], [15] donde se espera que los padres con ingresos familiares altos o aquellos que poseen un nivel educativo alto, tendrían mayores probabilidades de estudiar carreras con mayor duración.

Finalmente, en relación con las variables que afectan a la elección de carreras relacionadas a ciencias e ingeniería, un estudio realizado con un análisis Logit multinomial, concluyen que las siguientes variables afectan al escoger una carrera: tanto los estudiantes varones como las mujeres cuyos padres están en ocupaciones profesionales o ejecutivas tienen más probabilidades de optar por dichas áreas [12]. En este mismo sentido, existe otro estudio que explica el caso de la elección de un programa tecnológico, donde las variables que influyen negativamente es la educación de los hermanos y sus padres, el nivel de ingreso medio alto y tener computador [16].

Si bien los antecedentes expuestos evidencian de las variables que incidirían en la elección de una carrera en áreas generales como ingeniería son la situación socioeconómica, la educación de los padres y el rendimiento escolar, ninguno de ellos hace un contraste desde un punto de vista geográfico ni tampoco específico como carreras relacionadas a cierta materia o área. Por último, se ve una oportunidad para delimitar los estudios por zonas geográficas, ya que no siempre un país cuenta con características similares en términos de educación u oportunidades laborales, si es que uno, por ejemplo, contrasta la capital de un país con ciudades extremas a esta. Por ende, se espera que el trabajo realizado exponga variables similares, pero no iguales.

#### B. Minería de datos en la educación

La minería de datos en la educación (MDE), de acuerdo con [17] persigue descubrir patrones y hacer predicciones que caractericen las conductas y los logros de los estudiantes, el conocimiento de dominios contenidos, evaluaciones, entre otras. La creación de repositorios públicos de datos educacionales ha creado una base que hace posible la minería de datos. En particular, los datos de estos repositorios son totalmente válidos (ya que son datos reales sobre las características, rendimiento y aprendizaje de estudiantes), y cada vez más accesibles para comenzar una investigación. Algunos de los trabajos realizados en MDE sobre predicción son de un 60% y descriptivos un 40%, y además, destaca la creación de modelos predictivos para estimar la deserción y retención de los estudiantes [18].

La creación de modelos predictivos requiere tener aspectos a considerar para su desarrollo. Uno de estos es la metodología por emplear y una de ella es *Knowledge Discovery from Databases* (KDD). De hecho, gran mayoría de las metodologías nacen de esta última, ya que fue la primera en aceptarse como metodología para trabajar con los datos [19]. Dicha metodología propone cinco fases, de las cuales son la selección, preprocessamiento, transformación, minería de datos e interpretación. Gran parte del trabajo se invierte en las etapas de preprocessamiento y transformación, debido análisis exhaustivo y realizar una reducción de las variables a un conjunto óptimo para la etapa de minería de datos.

Otro aspecto que considerar es el tipo de técnicas de limpieza de datos, ya que es muy probable que exista falta de datos en una o más variables. La literatura propone una gama de técnicas, entre ellas la eliminación de registros completos que tengan alguna variable faltante [20]. Generalmente se recomienda la eliminación de registros cuando se posea una gran cantidad de valores perdidos o blancos y correspondan a una pequeña cantidad dentro del conjunto de datos [21], de no ser ese el caso es conveniente evaluar otras técnicas.

En relación con las características de los datos, las escasas ocurrencias de eventos perjudican la detección en modelos predictivos relacionados a la clasificación, donde este tipo de eventos recibe el nombre de desbalance de clases. Este desbalance hace referencia a un conjunto de datos dentro del cual una o algunas de las clases tienen un número mucho mayor de ejemplos que los demás. La clase prevalente recibe el nombre de clase mayoritaria, mientras que la clase poco común se denomina clase minoritaria [22]. Este tipo de comportamiento en los datos es más usual de lo que se cree, de hecho, aprender

de conjuntos de datos desequilibrados es uno de los 10 principales desafíos en la investigación de minería de datos. Varios investigadores han propuesto técnicas para trabajar bajo estas condiciones, una de ellas son las técnicas de re-muestreo, donde sus resultados han demostrado ser acordes a lo esperado [23].

Otro aspecto que es indispensable en los trabajos de minería de datos es la selección de variables. Los investigadores en años anteriores se percataron que, para tener una minería de datos exitosa, es necesario que la selección de variables sea un paso para considerar. De hecho, este proceso trae consigo aspectos positivos para el proceso de minería de datos, como es el aumento de velocidad de aprendizaje de los algoritmos, mejorar la precisión en los modelos y finalmente, ayudan a comprender mejor el modelo final [24]. Existen diversos métodos para la selección de variables relevantes, los que se destacan los métodos de filtro como la prueba de Chi-cuadrado, correlación de Pearson y ganancia de información. También están los métodos de envoltura y embebidos [25].

Luego de definir los aspectos anteriores, se procede a determinar las técnicas de aprendizaje para el modelamiento. Las técnicas aplicadas pueden ser de caracterización, asociación, clusterización, clasificación y regresión [26]. Las tareas relacionadas con la clasificación hacen referencia a la construcción de modelos donde se categorizan los registros en grupos predefinidos o clases ya conocidas. De hecho, la combinación de técnicas genera en muchos casos, un aumento en el rendimiento [27]. Estas técnicas de aprendizaje requieren que, para aprender, se divida el conjunto en entrenamiento y validación. Existen técnicas como la validación cruzada, que particiona el conjunto de datos en k partes mutuamente excluyentes, conteniendo un número similar de observaciones en cada partición [28].

Finalmente, los métodos y las técnicas descritas anteriormente deben ser implementadas en un entorno para ejecutarse. Actualmente existen diversas herramientas y softwares para la implementación de procesos relacionados con minería de datos, los cuales los más usados son, por ejemplo, RapidMiner [29]. El uso de RapidMiner se debe a su versatilidad y variedad de modelos y parámetros. Además de contar con una interfaz gráfica intuitiva y el sistema “Drag and drop”.

### III. METODOLOGIA DE TRABAJO

#### A. Selección

Esta etapa consistió en solicitar al Departamento de Evaluación, Medición y Registro Educacional (DEMRE) [30] su repositorio de información, la cual corresponde a la base de datos del proceso de admisión a la educación universitaria en Chile. El repositorio cuenta con algunas variables como el número de estudiantes inscritos para rendir la Prueba de Selección Universitaria (PSU) a nivel nacional y regional y por dependencia administrativa, antecedentes familiares y económicos del estudiante, los promedios y puntajes de las pruebas PSU. Además, cuenta con la variable género y la universidad a la cual se matriculó el alumno.

Al tener la base de datos, se filtró por el género (femenino) y que se hayan matriculado en una universidad del norte del

país perteneciente al CRUCH. Finalmente se creó una variable binaria, la cual indica si se matricula o no a una carrera relacionada con computación o informática, la que consta de toda carrera ya sea de ejecución o civil que tenga en su nombre informática o computación y corresponde a la variable dependiente. Esto se realizó con el objetivo de construir correctamente el conjunto de datos, ya que los datos estaban distribuidos en diversos archivos. El conjunto inicial de datos era de 41.000 registros aproximadamente, que comprendía los años 2004 hasta 2018 y con 210 variables.

#### B. Análisis previo

Antes de pasar al preprocesamiento como tal, se realizó un análisis descriptivo previo para obtener así una noción inicial del comportamiento de los datos. Algunos de los aspectos analizados fueron la cantidad de alumnas matriculadas en el norte clasificadas por región del establecimiento de egreso (escuela) y se aprecia en la Tabla I. También se exponen los porcentajes de las alumnas matriculadas en alguna universidad del norte según el tipo de establecimiento de egreso del cual provenían, el cual corresponde a la Fig. 1.

TABLA I. CANTIDAD DE ALUMNAS MATRICULADAS EN ALGUNA UNIVERSIDAD DEL NORTE SEGÚN REGION DEL ESTABLECIMIENTO DE EGRESO

Region del establecimiento de egreso	Mujeres matriculadas en otras carreras	Mujeres matriculadas en informática
Coquimbo	9.597	152
Antofagasta	8.641	105
Tarapacá	8.122	113
Arica	5.384	50
Atacáma	5.189	119
Metropolitana	1.083	14
Araucania	910	0
Valparaiso	542	1
BioBio	475	2
O'Higgins	378	3
Maule	173	2
Lagos	95	0
Aisen	28	0
Ríos	25	0
Magallanes	22	0

En la Tabla I se observó que la región de Coquimbo es de donde provienen más alumnas que se matriculan a una carrera relacionada con informática o computación en el norte, con 152 registros y también esta región lidera en aquellas que se matriculan a otras carreras, con 9.597 registros. La segunda región del establecimiento de egreso de donde provienen las alumnas que se matricularon al área de informática es la región de Atacáma, mientras que en aquellas que se matricularon en otras carreras es la región de Antofagasta. En resumen, se observa que no hay mucha presencia de alumnas que provengan de escuelas fuera de las regiones del norte, exceptuando la región Metropolitana para ambos casos.

Por otro lado, del total de mujeres matriculadas en Informática en el norte del país, resulta relevante conocer la procedencia según tipo de establecimiento para constatar si existen diferencias significativas. Esto apreciar en la Fig. 1.

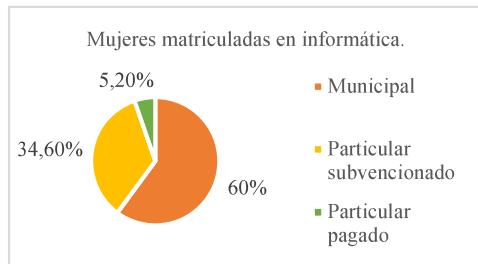


Fig. 1. Porcentajes de mujeres matriculadas en el norte por tipo de dependencia de egreso.

Para la Fig. 1, se seleccionaron todos los registros de aquellas que se hayan matriculado en universidades del norte y en informática y se analizó por tipo de dependencia del cual egresó. Se observó un gran número de mujeres que provenían de establecimientos municipales con un 60%, mientras que después proseguían los colegios particulares subvencionados. Estos datos resultan interesantes si es que se piensa en los procesos de admisión de las universidades.

### C. Preprocesamiento

Se aplicaron distintas técnicas de análisis de datos como el uso de estadística, visualización de datos, transformación de variables y reducción de los registros. En primer lugar, se analizó los tipos de variables, cantidades y distribución de las instancias. En el conjunto inicial se realizó un análisis exploratorio, donde se detectó un gran número de variables que no tenían relación al tema de la investigación, ya que estaban relacionadas con otros tipos de admisiones o que existían gran cantidad de datos faltantes en sus registros. Lo primero que se consideró es que cada universidad posee dos tipos de ingreso: regular y especiales, de los cuales solo se enfocó en el regular (rendición PSU), debido a que gran parte de las matrículas se enfocaban a este tipo de ingreso y porque los otros cupos restantes están orientados a otros tipos de estudiantes, siendo así requisitos totalmente distintos en comparación con los requerimientos de la admisión por PSU.

Otro criterio que se consideró fue la limpieza de los datos, el cual, según [21], era recomendable eliminar las variables y registros con datos faltantes siempre y cuando no fueran una cantidad importante en relación con el conjunto, por ende, se eliminaron aquellas variables con un porcentaje faltante de más de un 50% y se eliminó aquellos registros que presenten una o más variables con datos faltantes. Esto condujo a una cantidad de 40 variables y aproximadamente 32.000 registros a trabajar.

### D. Transformación

Posterior, para obtener un conjunto óptimo de datos, se procedió a la transformación de las variables, entre ellas la edad, puntajes PSU e ingreso bruto. La edad se calculó restando el año del proceso de admisión menos el año de nacimiento, esto con el fin de saber la edad que tuvieron las estudiantes cuando se matricularon. Por otra parte, los puntajes PSU se unificaron en 2 columnas, ya que algunas estudiantes postularon con puntajes anteriores y otras con puntajes actuales.

En relación con el ingreso bruto, este se encontraba por rangos, y cada cierto periodo cambiaban, por tanto, podrían contener inconsistencia. Una de las formas para mitigar esta

diferencia fue transformar los rangos de los ingresos en el valor del dinero actual, el cual considera la inflación a través de los años y con ello aplicar discretización a los rangos. Se utilizó la calculadora de IPC del Instituto Nacional de Estadística [31], donde permite calcular la tasa de variación del Índice de los Precios al Consumidor (IPC) entre dos años. Con el fin de adaptar los rangos salariales, se prosiguió en convertir los ingresos a las categorías de los grupos socioeconómicos en Chile, según la encuesta de Caracterización Socioeconómica (Casen), y la Encuesta de Presupuestos Familiares realizado por la Asociación de Investigadores de Mercado (AIM) [32], donde definen 7 grupos que son denominados: "E", "D", "C3", "C2", "C1b", "C1a", "AB". El resultado de todo este proceso se aprecia en la Tabla II.

TABLA II. PROCESO DE TRANSFORMACION DEL INGRESO BRUTO

Rango de ingresos	Estrato socioeconómico	Categoría
(0 hasta 324.000)	E	1
(324.0001 hasta 562.000)	D	2
(562.0001 hasta 899.000)	C3	3
(899.001 hasta 1.360.000)	C2	4
(1.360.000 o más)	ABC (C1b, C1a, AB)	5

Luego de la del proceso de la Tabla II, se procedió a la selección de las variables. Como se expuso en [25] sobre los métodos existentes, se seleccionó los métodos de filtro y se aplicó las pruebas de Chi-cuadrado y correlación de Pearson, debido a que existían variables tanto nominales y continuas y estas pruebas eran adecuadas para esos tipos de datos.

Posteriormente, se realizó el análisis de la distribución de la variable de salida. Al observar la Fig. 2, la clase uno son las mujeres que se matricularon a una carrera de informática y la clase cero son las que se matricularon a otras carreras.

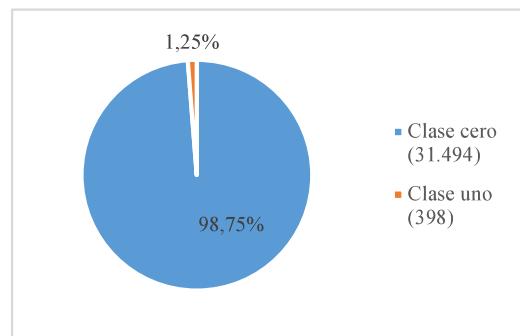


Fig. 2. Distribución de las clases de la variable dependiente.

Debido a que se observó un desbalance pronunciado, se emplearon las técnicas de submuestreo y sobremuestreo expuestas en el marco teórico. Para facilitar la comparación del rendimiento de los modelos, el conjunto de datos se separó según el siguiente esquema:

- Primer conjunto: Conjunto de datos sin aplicación de técnica de re-muestreo.
- Segundo conjunto: Conjunto con aplicación de técnica de balanceo "Submuestreo"

- Tercero conjunto: Conjunto con aplicación de técnica de balanceo “Sobremuestreo”.

#### E. Minería de datos.

Como se expuso en [26], existen diversas técnicas dependiendo del objetivo a investigar. Para este caso se definió como un problema de clasificación y se seleccionaron los siguientes algoritmos a trabajar en base a los modelos utilizados por trabajos similares [33], [34], Árbol de decisión, clasificador bayesiano, K-NN y bosques aleatorios. Por último, para implementar todo el proceso de selección de variables y el modelado se realizó con RapidMiner [35].

Como se emplearon tres conjuntos de datos, se debían configurar diferentes parámetros en cada uno de los modelos, pero esta tarea consumía excesivos recursos de hardware para encontrar aquellos que fueran los más adecuados a cada modelo. Por esta razón, se estableció que los parámetros de los modelos se ingresarían según el operador de optimización de parámetros que se encuentra en RapidMiner, el cual facilita la búsqueda de un conjunto óptimo de parámetros para los operadores utilizados.

La validación de los modelos se realizó por medio de la técnica de *cross-validation* y también cuando se trabaja con los resultados de los modelos, se deben indicar las métricas a evaluar. Se estableció como medida central el *Area Under Curve* (AUC) y también se consideró las siguientes métricas: predicción, sensibilidad y especificidad.

#### F. Interpretación y evaluación

Con el fin de cuantificar la importancia de las variables, se seleccionó el criterio de ganancia de información (GI), el cual cuantifica la importancia según la clase a predecir. La selección de este método se basó en lo expuesto en los antecedentes presentado por [27], [36], y además este criterio no presenta restricciones dentro de RapidMiner sobre el tipo de variable a analizar. Asimismo, el conjunto de datos cumplía que la variable a predecir fuera de tipo nominal [37].

GI calcula cuánta información aporta cada variable respecto de la clase a predecir utilizando el cálculo de la entropía. La entropía denota el nivel de incertidumbre que presenta una distribución, cuanta más entropía más incertidumbre [36]. Se destaca que GI es cero si y solo si las variables son independientes, es decir no comparten ninguna información entre sí. No obstante, GI presenta inconvenientes cuando los datos de las variables toman un gran rango de valores únicos. Por ejemplo, una variable que sea el número de tarjeta de crédito de un cliente puede tener una alta ganancia de información. Esto sucede porque identifica de forma única a cada cliente, pero es posible que no se desee asignar una gran ponderación a dicha variable [38]. Para el caso del conjunto de este trabajo, las variables analizadas no presentan este tipo de inconvenientes. Por tal motivo, se utilizó RapidMiner ya que incluye un operador que realiza el cálculo del peso de la variable en relación con este criterio. Ya que GI establece que los datos deben estar discretizados, este operador prueba todos los puntos de división posibles entre dos números vecinos, luego selecciona el punto de división con la ganancia más alta y entrega el valor correspondiente [39]. En la Tabla III se presentan las variables del conjunto de datos ordenadas descendente según su

peso. Al observar la Tabla III, da cuenta de un orden de las variables que coincide en gran medida con lo propuesto por la literatura, donde el rendimiento de la estudiante está presente en las primeras cinco variables, como también ubicación geográfica del establecimiento y la educación de la madre. Es importante destacar que la literatura proponía que la variable relacionada a las notas o puntajes de la prueba de matemáticas era una de las más relevantes, mientras que en esta investigación la prueba de PSU de lenguaje está siendo más relevante.

TABLA III. VARIABLES MAS RELEVANTES SEGÚN PESO

Variable	Definición	Peso
Reg_Est	Región del establecimiento educacional	1.00
Ptje_Len	Puntaje PSU de lenguaje	0.91
Prom	Promedio de notas de enseñanza media	0.81
Edu_M	Educación de la madre	0.50
Ptje_Mat	Puntaje PSU de matemáticas	0.39
Ram_Edu	Rama educacional del establecimiento	0.36
Finan_P	Tipo de financiamiento para la universidad	0.34
Vivi_Con	Con quien vivirá al entrar a la universidad	0.22
Edu_P	Educación del padre	0.20
Jefe_H	Jefe del hogar	0.20
N_Grupo	Nº de personas que conforman el grupo familiar	0.16
N_E_Eds	Nº de personas del grupo que estudian ed. superior	0.08
N_E_I_III	Nº de personas del grupo que estudian I a III medio	0.04
N_G_Tra	Nº de personas del grupo que trabajan	0.01
N_E_Otr	Nº de personas que estudian otras	0.01
N_E_IV	Nº de personas que estudian IV medio	0.00

A continuación, en las Tablas IV, V y VI se presentan los resultados de rendimiento que obtuvieron los modelos para cada uno de los tres conjuntos de datos definidos.

TABLA IV. COMPARACIÓN DEL RENDIMIENTOS DE LOS MODELOS CON EL PRIMER CONJUNTO.

Modelo	Predicción	Sensibilidad	Especificidad	AUC
Clasificador bayesiano	98,71%	0,0%	99,96%	0,65
Bosques aleatorios	97,65%	2,98%	98,84%	0,51
K-NN	97,55%	4,03%	98,73%	0,50
Árbol de decisión	97,22%	2,01%	98,42%	0,50

TABLA V. COMPARACIÓN DEL RENDIMIENTOS DE LOS MODELOS CON EL SEGUNDO CONJUNTO.

Modelo	Predicción	Sensibilidad	Especificidad	AUC
Clasificador bayesiano	60,82%	61,37%	60,82%	0,63
Bosques aleatorios	60,24%	57,14%	60,28%	0,62
K-NN	59,96%	53,08%	60,05%	0,58
Árbol de decisión	56,05%	55,05%	56,07%	0,46

TABLA VI. COMPARACIÓN DEL RENDIMIENTOS DE LOS MODELOS CON EL TERCER CONJUNTO.

Modelo	Predicción	Sensibilidad	Especificidad	AUC
Clasificador bayesiano	58,87%	54,74%	58,92%	0,60
Bosques aleatorios	68,68%	40,90%	69,03%	0,56
K-NN	61,44%	47,24%	61,62%	0,56
Árbol de decisión	52,46%	52,63%	52,45%	0,52

En primer lugar, para evaluar los resultados, se establece el orden a considerar de las métricas según lo mencionado al final de la sección minería de datos.

El rendimiento del primer conjunto da origen a una gran desequilibrio en las métricas de sensibilidad y especificidad, como se esperaba por la literatura. En la Tabla IV se observa que los porcentajes de predicción son elevados, más de un 90%, aunque si se compara con las otras métricas, como sensibilidad o AUC, están son bajas. AUC tiene la característica que a partir del valor 0.5, ascienda su probabilidad de predecir ambas instancias correctamente, y el único modelo que alcanzó un AUC mayor a 0.5 fue el clasificador bayesiano, pero al observar la métrica especificidad, deja en evidencia que solo está clasificando una clase solamente. Considerando que ningún modelo clasificó exitosamente más de un 4% respecto a la sensibilidad, se descartan estos resultados.

Por otra parte, al analizar la Tabla V, se aprecia que los resultados de las métricas están mejor balanceados si se compara con la Tabla IV. El Clasificador bayesiano obtuvo un AUC de 0.63, el mayor de toda la Tabla V y una predicción aproximada de un 61%, mientras que la sensibilidad y especificidad están equilibradas en comparación a los resultados anteriores de la Tabla IV. Por otra parte, Bosques Aleatorios es el que obtiene una predicción y un valor AUC muy cercano al modelo bayesiano, aunque si se observa la Tabla V, el porcentaje de sensibilidad es menor al modelo bayesiano, de igual manera la métrica especificidad, aunque esta diferencia es menor. K-NN obtuvo un valor de predicción de un 59%, aunque este modelo acierta en más instancias en la especificidad y su valor AUC es aceptable, pero si se compara con otros modelos de la misma tabla, obtuvo un rendimiento medio. Finalmente, árbol de decisión obtuvo un 55% y un 56% en sensibilidad y especificidad, aunque el valor de AUC es bajo, un 0.46. Bosques aleatorios y el Clasificador bayesiano estuvieron muy cercanos en los resultados, pero el modelo bayesiano obtuvo más porcentaje en sensibilidad que Bosques aleatorios, por ende, de la Tabla V el modelo que tiene métricas mejor balanceadas es el Clasificador bayesiano.

En la Tabla VI, nuevamente el Clasificador bayesiano obtuvo el mayor AUC dentro de toda la tabla, aunque si se observa la métrica sensibilidad logró un menor porcentaje que la especificidad. Luego al observar el rendimiento de bosques aleatorios donde la métrica más alta es en especificidad, con un 69%, lo que indica que identificó bien a la clase cero. Por el contrario, este modelo obtuvo el porcentaje más bajo de sensibilidad, lo que impacta también en su aumento en el

porcentaje de predicción y en consecuencia, su AUC es solo de un 0.56. K-NN obtuvo una predicción medianamente alta, que solo se ve superada por bosques aleatorios, aunque su porcentaje de sensibilidad es bajo, solo un 47%, mientras que la especificidad es de un 61%. El modelo de árbol de decisión obtuvo un rendimiento menor que los otros clasificadores de la Tabla VI, debido a que obtuvo un AUC bajo y sus porcentajes de sensibilidad y especificidad lo son también. De todos los modelos de la Tabla VI, el Clasificador bayesiano fue quien obtuvo las mejores métricas, pero si se compara con el modelo bayesiano de la Tabla V, este obtuvo mejores métricas, por ende, el modelo bayesiano del segundo conjunto es escogido.

#### IV. DISCUSIÓN Y CONCLUSIONES

La presente investigación tuvo por objetivo identificar variables que caractericen a los estudiantes que ingresaron a una carrera relacionada con informática o computación en universidades pertenecientes al CRUCH del norte de Chile. Esto se logra mediante la selección de variables relevantes, aplicación de técnicas de balanceo de clases, selección de técnicas de aprendizaje y finalmente, la comparación del rendimiento de los modelos mediante una tabla de métricas. Al considerar todas las variables relevantes por el criterio de ganancia de información y compararlas con lo visto en el marco teórico, se observó que existe un patrón en la relevancia de las variables a la hora de elegir los predictores que influyen en la elección de una carrera. Un aspecto que llama la atención de este trabajo fue que el criterio de GI asignó mayor valor al puntaje PSU de lenguaje, ante que el puntaje de matemática. Según la literatura, varios estudios mencionan la importancia de las variables relacionadas con el aspecto matemático para optar a carreras de ingeniería, pero ninguna habla de aquellas que tengan que ver con el aspecto del lenguaje. Esto podría suponer una oportunidad para indagar más, ya que generalmente las admisiones de las Universidades Chilenas restan importancia a esta variable y les asignan a otras como los puntajes PSU de matemáticas y ciencias.

Por otra parte, hubo variables que concordaron con la literatura, como lo fue en el caso del rendimiento escolar, la educación de los padres y el tipo de financiamiento de la educación superior. Estas variables se agruparon dentro de las primeras diez, siendo la educación del padre la penúltima variable. Es interesante y sería motivo de estudio posteriores, analizar en más profundidad el impacto de la educación de los padres en la elección de carreras por parte de las alumnas.

Finalmente, hubo una variable que no estuvo dentro de la Tabla III y que la literatura mencionaba que era relevante, que fue el ingreso económico. Esto pudo haber sucedido por cómo estaba la variable en un principio y se tuvo que recurrir a trasformaciones para poder trabajar con ella. Sería útil encontrar otro tipo de tratamiento a esta variable para rescatar toda la información posible.

Respecto al rendimiento del primer conjunto, el desbalanceo de clases provocó resultados que solo tendían a clasificar la clase cero. Esto se observa por su alto porcentaje de especificidad y bajo en la sensibilidad, además que este efecto era probable debido a que la literatura lo mencionaba. Por esto, no se consideró este conjunto ya que el foco estaba en la clasificación correcta de ambas clases.

En el segundo y tercer conjunto, se observó una mejoría en las métricas por parte de los modelos a identificar la clase minoritaria, aunque esta era, en algunos casos, el mismo porcentaje de aciertos que la clase mayoritaria. Con la aplicación de la técnica del submuestreo, se observó que los modelos no estaban sesgados. En sí, los resultados estuvieron bastante cercanos unos de otros, excepto en algunos casos como lo fue en los porcentajes de predicción y sensibilidad. El Clasificador bayesiano y Bosques aleatorios en el segundo conjunto tuvieron AUC muy cercanos, pero su diferencia fue en la métrica de sensibilidad, haciendo que el Clasificador bayesiano clasificara mejor las instancias de la clase uno. Este modelo de entre todos fue el que obtuvo métricas mejor balanceadas con la aplicación de la técnica de submuestreo con un 61% de predicción y AUC de 0.63. Si bien se escogieron cuatro modelos de clasificación y técnicas de re-muestreo, estos obtuvieron rendimientos que podrían eventualmente mejorarse. Aspectos como un tratamiento de los datos más profundo como la transformación de las variables en otras clases, redistribución, aplicación de otras técnicas de balance de clases y de técnicas de limpieza más sofisticadas, podrían representar una opción válida para mejorar los resultados. Los trabajos existentes proponen que, al tener comportamientos muy similares o se desee mejorar los resultados de los modelos, se opten por modelos combinados, el cual une técnicas para mejorar la detección de patrones y rendimiento. También está el factor de la elección de los modelos, donde la evaluación de otros algoritmos sería conveniente a la hora de poder mejorar las métricas de evaluación, si se desea implementar un modelo que clasifique con mayor precisión.

Este trabajo presenta antecedentes que pueden ser utilizados por las unidades de las escuelas y establecimientos de educación superior del norte de Chile para enfocar aún más las actividades de admisión. También esta investigación expone información acerca de las variables relevantes para que las estudiantes puedan tener un mayor conocimiento acerca de la elección de una carrera profesional.

Finalmente, este tipo de estudios podría eventualmente replicarse para las diferentes carreras de STEM y áreas geográficas, ya que estas están siendo requeridas debido a la transformación digital y también hay que considerar que no todas las regiones o estados de un país poseen características y necesidades idénticas.

#### AGRADECIMIENTOS

Se agradece al Departamento de Evaluación, Medición y Registro Educativo (DEMRE), de la Universidad de Chile, por facilitar las bases de datos del Sistema de Admisión a la Educación Superior Universitaria para el desarrollo de esta investigación.

#### REFERENCES

- [1] UNESCO, *Cracking the code: girls' and women's education in science, technology, engineering and mathematics (STEM)*, Revised ve. 2017.
- [2] OCDE, "The Pursuit of Gender Equality," Oct. 04, 2017. [https://www.oecd-ilibrary.org/social-issues-migration-health/the-pursuit-of-gender-equality\\_9789264281318-en](https://www.oecd-ilibrary.org/social-issues-migration-health/the-pursuit-of-gender-equality_9789264281318-en) (accessed Jun. 28, 2020).
- [3] F. Fernández and M. Márquez, "Participación de mujeres en carreras informáticas en chile 1: ¿quienes eligen informática hoy?", 2019. <http://www.opinionsocial.cl/wp-content/uploads/2019/03/PARTICIPACION-DE-MUJERES-EN-CARRERAS-INFORMATICAS-EN-CHILE.docx.pdf> (accessed Jul. 23, 2020).
- [4] B. J. Drury, J. O. Siy, and S. Cheryan, "When Do Female Role Models Benefit Women? The Importance of Differentiating Recruitment From Retention in STEM," *Psychol. Inq.*, vol. 22, no. 4, pp. 265–269, 2011, doi: 10.1080/1047840X.2011.620935.
- [5] K. S. Dubina, T. L. Morisi, M. Rieley, and A. B. Wagoner, "Projections overview and highlights, 2018–28," *Mon. Lab. Rev.*, vol. 142, p. 1, 2019.
- [6] E. Mannix and M. A. Neale, "What differences make a difference? The promise and reality of diverse teams in organizations," *Psychol. Sci. public Interes.*, vol. 6, no. 2, pp. 31–55, 2005.
- [7] B. O. Camarena Gómez, D. González Lomelí, and D. Velarde Hernández, "El programa de orientación educativa en bachillerato como mediador en la elección de carrera," *Rev. Mex. Investig. Educ.*, vol. 14, no. 41, pp. 539–562, 2009.
- [8] J. S. Eccles, "Understanding women's educational and occupational choices," *Psychol. Women Q.*, vol. 18, no. 4, pp. 585–609, 1994, doi: 10.1111/j.1471-6402.1994.tb01049.x.
- [9] G. Bravo and M. Vergara, "Factores que determinan la elección de carrera profesional: en estudiantes de undécimo grado de colegios públicos y privados de Barrancabermeja," *Psicoespacios*, vol. 12, no. 20, pp. 35–48, 2018.
- [10] C. Celestino and M. Alicia, "Motivación y elección de carrera," *Rev. Mex. orientación Educ.*, vol. 5, no. 13, pp. 6–9, 2008.
- [11] E. Gámez and H. Marrero, "Metas y motivos en la elección de la carrera universitaria: Un estudio comparativo entre psicología, derecho y biología," *An. Psicol. Psychol.*, vol. 19, no. 1, pp. 121–131, 2003.
- [12] K. Leppel, M. L. Williams, and C. Waldauer, "The impact of parental occupation and socioeconomic status on choice of college major," *J. Fam. Econ. Issues*, no. 22, pp. 373–394, 2001, doi: 10.1023/A:1012716828901.
- [13] S. R. Porter and P. D. Umbach, "College major choice: An analysis of person–environment fit," *Res. High. Educ.*, vol. 47, no. 4, pp. 429–449, 2006.
- [14] K. Leppel, M. L. Williams, and C. Waldauer, "The impact of parental occupation and socioeconomic status on choice of college major," *J. Fam. Econ. Issues*, vol. 22, no. 4, pp. 373–394, 2001, doi: 10.1023/A:1012716828901.
- [15] L. Di Gresia, "Educación universitaria: acceso, elección de carrera y rendimiento," Universidad Nacional de La Plata, 2009.
- [16] J. Miranda Morales, H. Gutiérrez, J. Manotas, and V. Higuera, "Evidencia empírica sobre la teoría de la demanda de educación superior en América Latina: un estudio sobre el caso del Caribe Colombiano," pp. 19–37, 2014, doi: 10.13140/2.1.3773.0089.
- [17] J. Luan, "Data Mining and Its Applications in Higher Education," *New Dir. Institutional Res.*, vol. 2002, no. 113, pp. 17–36, 2002, doi:

- 10.1002/ir.35.
- [18] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.
- [19] G. Mariscal, O. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *Knowl. Eng. Rev.*, vol. 25, pp. 137–166, 2010, doi: 10.1017/S0269888910000032.
- [20] J. Scheffer, "Dealing with missing data," *Res. Lett. Inf. Math. Sci.*, vol. 3, pp. 153–160, 2002.
- [21] E. Baldizzoni, "Propuesta de Proceso de Transformación de Datos para Proyectos de Explotación de Información," *Rev. Latinoam. Ing. Softw.*, vol. 1, p. 56, 2014, doi: 10.18294/relais.2013.56-70.
- [22] A. Shillabeer and K. Jackson, "Gender Imbalance in Undergraduate IT Programs – A Vietnamese Perspective," *Innov. Teach. Learn. Inf. Comput. Sci.*, vol. 12, no. 1, pp. 70–83, 2013, doi: 10.11120/ital.2013.00005.
- [23] R. I. Rashu, N. Haq, and R. M. Rahman, "Data mining approaches to predict final grade by overcoming class imbalance problem," in *2014 17th International Conference on Computer and Information Technology (ICCIT)*, 2014, pp. 14–19, doi: 10.1109/ICCITECHN.2014.7073095.
- [24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003, [Online]. Available: <http://www.jmlr.org/papers/v3/guyon03a.html>.
- [25] Á. Flores Ríos, "Métodos de selección de atributos basados en utilidades para la predicción de fuga de clientes en telecomunicaciones," Universidad de Chile, 2014.
- [26] M. K. Jiawei Han and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, Amsterdam, 2011.
- [27] F. Bugueño, "Modelo predictivo para la selección de postulantes destacados a una institución de educación superior," Universidad de Chile, 2017.
- [28] Y. Liu, "The evaluation of classification models for credit scoring," *Institut für Wirtschaftsinformatik, Georg-August-Universität Göttingen*, 2002. <http://webdoc.sub.gwdg.de/ebook/lm/arbeitsberichte/2002/02.pdf>
- [29] H. M. Marin-Castro and P. E. Franco-Vázquez, "Estudio de Herramientas de Minería de Datos para la Tarea de Clasificación," *TecnoINTELECTO*, vol. 14, no. 1, pp. 1–9, 2017, [Online]. Available: [http://www.itvictoria.edu.mx/personal/tecnoinlecto/TecnoINTELLECTO- Vol 14\(1\)-SEP-29-2017.pdf#page=3](http://www.itvictoria.edu.mx/personal/tecnoinlecto/TecnoINTELLECTO- Vol 14(1)-SEP-29-2017.pdf#page=3).
- [30] DEMRE, "Prueba de Transición para la Admisión Universitaria - DEMRE." <https://demre.cl/> (accessed Jul. 13, 2020).
- [31] INE, ...:"Calculadora IPC:..." <https://calculadora.ine.cl/> (accessed Jul. 13, 2020).
- [32] AIM, "AIM – Asociación de Investigadores de Mercado." <https://www.aimchile.cl/> (accessed Jul. 13, 2020).
- [33] J. Díaz-Ramírez, X. Badilla-Torrico, and J. Martí-Lara, "Data mining: A methodological proposal for higher education," *Opcion*, vol. 35, no. SpecialEdition25, 2019, [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-85085900145&partnerID=MN8TOARS>.
- [34] J. P. DEL RÍO ARTEAGA, "Modelo predictivo para la retención de estudiantes en primeros años," 2018.
- [35] RapidMiner, "RapidMiner | Best Data Science & Machine Learning Platform." <https://rapidminer.com/> (accessed Jul. 13, 2020).
- [36] J. V. Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature Selection with High-Dimensional Imbalanced Data," in *2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 507–514, doi: 10.1109/ICDMW.2009.35.
- [37] V. Kotu and B. Deshpande, *Predictive analytics and data mining: concepts and practice with RapidMiner*. Morgan Kaufmann, 2014.
- [38] RapidMiner, "Weight by Information Gain." [https://docs.rapidminer.com/latest/studio/operators/modeling/feature\\_weights/weight\\_by\\_information\\_gain.html](https://docs.rapidminer.com/latest/studio/operators/modeling/feature_weights/weight_by_information_gain.html) (accessed Oct. 12, 2020).
- [39] "Information gain and numerical attributes — RapidMiner Community." <https://community.rapidminer.com/discussion/258/information-gain-and-numerical-attributes> (accessed Oct. 12, 2020).
- (accessed Jul. 23, 2020).