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

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING PARA EL ESTUDIO DE DESERCIÓN TEMPRANA Y EGRESO OPORTUNO EN ESTUDIANTES DE INGENIERÍA DE LA FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

ROBINSON IGNACIO CASTRO LÓPEZ

PROFESOR GUÍA:
RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:
SERGIO CELIS GUZMÁN
ASTRID CONTRERAS FUENTES

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: ROBINSON IGNACIO CASTRO LÓPEZ
FECHA: 2020
PROF. GUÍA: RICHARD WEBER HAAS

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING PARA EL ESTUDIO DE DESERCIÓN TEMPRANA Y EGRESO OPORTUNO EN ESTUDIANTES DE INGENIERÍA DE LA FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

Los fenómenos de deserción universitaria y retraso académico significan un costo adicional al ya elevado costo que conlleva el paso por el sistema educacional. Estudios sugieren que estos fenómenos pueden ser predichos de manera anticipada al utilizar técnicas de machine learning.

El presente trabajo de título se hace partícipe del proyecto 'Desarrollo de tecnologías de Big Data para aumentar la retención y éxito de estudiantes universitarios' financiado por FONDEF (Fondo de Fomento al Desarrollo Científico), el cual se encuentra en una etapa de validación de técnica para el desarrollo de un prototipo que permita apoyar las labores de los gestores de permanencia en los establecimientos educacionales.

La investigación hace uso de una base de datos con 10.413 estudiantes de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. De esta forma, se suma a los estudios de otras universidades con los que cuenta el proyecto de manera previa, estos se realizaron en la Univerisdad Adolfo Ibañez, Univerisdad de Talca y Univerisdad Autónoma, sirviendo como marco de referencia para la investigación.

El trabajo realizado analiza distintos modelos de clasificación proveniente de una cartera de modelos, con el fin de identificar si existe alguno que muestre mejores rendimientos al explicar los fenómenos de estudio, y de esta forma ser recomendado para el desarrollo de dicha herramienta. Además, se identifican atributos que permiten un mejor rendimiento de los modelos de clasificación, siendo considerados estos como atributos relevantes para explicar los fenómenos.

Los resultados revelan que los atributos socio-demográficos son los más relevantes para explicar el fenómeno de deserción temprana, mientras que el promedio de enseñanza media es el atributo más importante para explicar el egreso oportuno.

Se recomienda para instituciones con características similares a la Facultad de Ciencias Físicas y Matemáticas el uso de los modelos Logistic Regresion, Neural Net y Super Vector Machine para explicar el fenómeno de deserción al primer año. Los modelos Naive Bayes y Logistic Regresion para estudiar el fenómeno de deserción al segundo año. Mientras que, se recomienda el modelo de Neural Net para estudiar el egreso oportuno.

El estudio utiliza diferentes métricas para evaluar los rendimientos de los modelos, mostrando que estos se desempeñan en niveles distintos según la dimensión de rendimiento. Por lo tanto, según los intereses de futuras investigaciones se recomienda utilizar métricas de rendimiento combinadas para la construcción de los clasificadores.

Con mucho cariño para mi mamá, papá y hermanos. Los quiero.

Tabla de Contenido

Capítulo 1

Introducción

1.1. Motivación

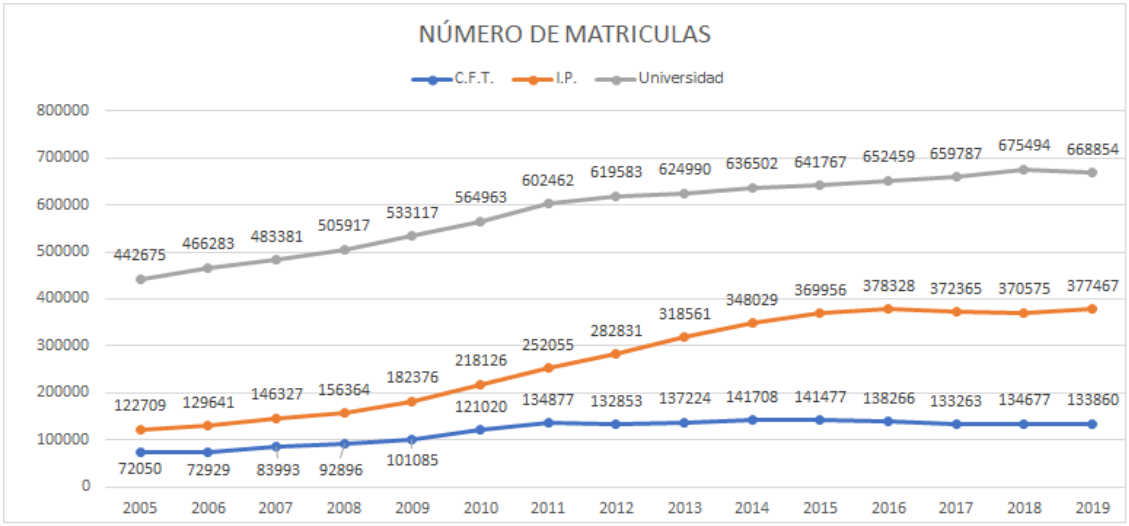
El sistema educacional superior chileno se vio fuertemente afectado por la reforma a principios de los años 80, al entrar en vigencia la Ley General de Universidades, dando inicio a la privatización de la enseñanza superior, creando establecimientos sin dependencia estatal e impulsando una masificación en la oferta de la educación superior.

Esto permitió la creación de nuevas instituciones de educación superior, entre estas, universidades, institutos profesionales y centros de formación técnica, al mismo tiempo disminuyó las barreras de entrada, permitiendo un aumento sostenido en el número de matrículas.

Además, se reestructuran las dos universidades estatales que contaban con presencia en diferentes regiones, forzándolas a desarticular sus sedes a lo largo del país, las cuales mediante diferentes recursos organizacionales dan lugar a la creación de 14 nuevas instituciones estatales. A esto se suma la creación de nuevos centros privados de estudios logrando alcanzar una cifra de 302 establecimientos de educación superior a principio de los años 90. Así, para año 2019 el sistema de educación superior en Chile se encuentra conformada por 159 instituciones, entre estas se hayan 60 universidades, 43 instituciones profesionales y 46 centros de formación técnica.

El gran nivel de cobertura ha impactado de manera positiva en el número de matrículas anuales, permitiendo que los distintos tipos de establecimientos educacionales muestren una tendencia, en general, creciente en el número de matrículas. Como se refleja en el gráfico de la figura 1.1.

Figura 1.1: Número de matrículas anuales en educación superior (elaboración propia, datos del CNE)



Hoy se considera que la educación superior es de carácter universal, permitiendo que distintos grupos sociales puedan ingresar a esta, y por lo tanto, deja de ser exclusivo para las clases más privilegiadas, generando mayores niveles de equidad. Sin embargo, el aumento en el número de matrículas y cobertura no deben ser los únicos indicadores al respecto, es necesario tener una visión más amplia para evaluar las políticas sobre la educación superior. Se debe buscar un sistema de educación que entregue calidad a sus estudiantes, este debe ser eficaz en el logro de los objetivos académicos para todos sus estudiantes, y eficiente en uso de los recursos con el fin de que exista una coherencia entre los beneficios que trae estudiar una carrera y los costos asociados que conlleva.

Con el aumento en el número de estudiantes que ingresan todos los años, se presenta una preocupación creciente por el desarrollo exitoso de sus carreras, tornándose relevante considerar fenómenos tales como la tasa de retención y el aumento de la duración real de las carreras, ya que el descuido de estos fenómenos puede provocar costos económicos para el estudiante sumada a una sensación de frustración, los que verían cómo sus carreras duran más de lo previsto e incluso peligran de ser finalizadas con éxito.

Estos fenómenos son estudiados anualmente por el Ministerio de Educación por medio de informes a cargo del Servicio de Información de Educación Superior (SIES) perteneciente a esta misma institución. Para entender la magnitud de los fenómenos señalados, se exponen los resultados correspondientes al estudio realizado para el año 2019. De esta forma el SIES define la retención estudiantil de primer año como el porcentaje de estudiantes que se mantiene en el mismo plan de estudios al año siguiente de realizar su ingreso. Los resultados muestran que, a nivel general, la tasa de retención alcanza un 75 % en los estudiantes de pregrado pertenecientes al cohorte del 2018[1]. Esto implica un 25 % de estudiantes que deserta de su plan de estudios [1].

[1] 70.5 % para Centros de Formación Técnica, 72.4 % para Institutos Profesionales y 78.9 % para Universidades

Por otra parte, el tiempo que los estudiantes se demoran en finalizar sus estudios superiores es un factor importante al momento de hacer la elección sobre el programa al que ingresan. Si bien, las instituciones educacionales informan sobre una duración formal que poseen las carreras, la realidad se contrapone de manera sustancial, mostrando una diferencia no menor entre la duración real y la duración formal. Esta sobreduración hace referencia a los semestres adicionales que le toma al estudiante egresar con respecto a la duración formal que informa el establecimiento. Alcanzando incluso una cifra de 31.5 % de sobreduración promedio para los distintos tipos de establecimientos. [2]

Estos datos revelan una realidad presente dentro de los estudiantes de educación superior en Chile y cuantifican fenómenos que dificultan su desarrollo exitoso. Si bien no se puede comprender en su totalidad el impacto social y económico que conllevan, si existen datos que revelan los costos asociados al ingreso a la educación superior para la población.

Con respecto a los costos económicos, es posible identificar un costo personal asociado a la inversión monetaria que conlleva el paso por el sistema universitario. En Chile se ve como los costos asociados al arancel corresponden a 27.9 % y 32 % del ingreso nacional bruto per cápita, para universidades públicas y privadas respectivamente. Así, se encuentran importantes diferencias cuando se compara Chile con otros países de la OCDE, en particular al analizar el costo de estudiar en una universidad pública. [3]

País	**U. Publicas**	**U. Privadas**
Australia	11.3 %	21.9 %
Canada	10.0 %	n/a
Japón	11.8 %	18.5 %
Corea	16.3 %	31.1 %
Nueva Zelanda	6.5 %	n/a
Reino Unido	5.2 %	4.9 %
Estados Unidos	11.4 %	42.0 %
Italia	3.3 %	11.5 %
Holanda	4.4 %	4.4 %
Israel	12.0 %	29.2 %
Chile	27.9 %	32.0 %

Tabla 1.1: Aranceles como porcentaje del ingreso nacional bruto per cápita

Si bien, existe un apoyo económico por parte del Estado por medio de becas y créditos universitario, estos valores reflejan el importante costo que deben incurrir las familias que no reciben apoyo para poder financiar sus estudios, teniendo un mayor impacto en los grupos familiares con ingresos medio-bajo, quienes deben endeudarse en cifras importantes para ingresar a un sistema que no garantiza el éxito del mismo, privatizando el fracaso del estudiante e incurriendo de manera directa los costos asociados a la deserción [4] o la sobre extensión de la carrera.

El paso por la educación superior se encuentra fuertemente relacionado con la percepción de mayores ingresos dentro del mercado laboral, permitiendo generar movilidad social, donde el 64.4 % de los estudiantes que ingresan al sistema corresponden a estudiantes de primera

generación, es decir, estudiantes cuyos padres no poseen un título universitario, sin embargo estos presentan un mayor riesgo de deserción con casi 3 veces más probabilidades de desertar que sus compañeros[5].

Considerando el crecimiento mantenido en el número matrículas en la educación superior, se estima que la deserción afectará a lo menos a unos 45,000 estudiantes por año, siendo la población más vulnerable la más afectada, impactando de manera directa a las familias y su sustentabilidad financiera. Por lo tanto se hace de gran relevancia el desarrollo de estrategias que residan en buscar una equidad de oportunidades para los distintos tipos de estudiantes [6].

1.2. Fondo Estatal FONDEF

El trabajo a realizar se enmarca en la adjudicación del fondo estatal FONDEF (Fondo al Desarrollo Científico y Tecnológico). Fondo que es utilizado con el fin de contribuir al mejoramiento en la calidad de vida de los chilenos por medio de la vinculación de instituciones investigadoras y empresas, aportando a la investigación y a el desarrollo tecnológico en los distintos sectores de intereses públicos.

Entre sus líneas de acción se encuentra el apoyo a proyectos de investigación y desarrollo aplicado con alto nivel de contenido científico, con el objetivo de generar un cambio positivo y significativo bajo un enfoque económico y/o social.

El presente trabajo se hace parte del desarrollo del proyecto 'Desarrollo de tecnologías de Big Data para aumentar la retención y éxito de estudiantes universitarios' financiado por medio del fondo estatal FONDEF, el cual busca el desarrollo de una herramienta genérica que logre identificar estudiantes propensos a desertar de manera temprana y estudiantes que finalizarán sus estudios de manera oportuna por medio de sus características especificas, con el fin de ser una herramienta de utilidad para los gestores de permanencia de los establecimientos de educación superior.

Actualmente el proyecto se encuentra en una etapa de validación de herramientas, en la que busca validar la utilización de modelos predictivos de minería de datos. La investigación hace participe a cuatro distintas casas de estudios, Universidad Adolfo Ibañez, Universidad Autónoma, Universidad de Talca y Universidad de Chile. Se definen tres fenómenos de interés para el estudio: deserción al primer año, deserción al segundo año y egreso oportuno.

El proyecto contempla el estudio para las cuatro universidades, de las cuales 3 ya se encuentran con resultados de investigación, faltando la investigación en los estudiantes de la Universidad de Chile, del cual se hace cargo el presente proyecto de título. Los resultados obtenidos en las otras universidades servirán como marco de referencia y de comparación.

De esta forma el proyecto busca aplicar técnicas de minería de datos qué permitan estudiar los fenómenos de deserción y egreso oportuno por medio de atributos socio-demográficos, atributos socio-económico, atributos académicos y atributos propios del proceso de selección universitaria, la investigación se realiza con datos de estudiantes de ingeniería de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

1.3. Justificación del Problema

La minería de datos forma parte de las ciencias de la computación y ha sido utilizada ampliamente en distintas industrias como una herramienta de apoyo en la toma de decisiones, puesto que permite descubrir conocimiento a partir de un conjunto de datos, encontrando patrones válidos, comprensibles y potencialmente útiles.

La aplicación de estas técnicas en estudios de fenómenos de la educación no es un tópico nuevo y ha sido muy relevante en los últimos años, ofreciendo herramientas que permiten predecir fenómenos de interés por medio de modelos de clasificación y/o de regresión. Esto, debido a que existe evidencia que correlaciona los atributos característicos y el comportamiento académico de los estudiantes.

En el caso chileno, se ve cómo los datos levantados por el Ministerio de Educación van en esta dirección[2]. Si bien, estos estudios constatan los fenómenos de retención estudiantil y sobre-duración de los programas educacionales, tales fenómenos se encuentran fuertemente relacionados a la deserción y al término oportuno del plan de estudio.

A continuación se muestran una serie de datos que permiten presumir una correlación entre el comportamiento académico y los atributos de los estudiantes. Estos se han obtenido desde el Servicio de Información de Educación Superior perteneciente al Ministerio de Educación.

Sexo	**2014**	**2015**	**2016**	**2017**	**2018**
Mujeres	73.30 %	74.20 %	75.60 %	76.50 %	77.40 %
Hombres	67.70 %	68.10 %	69.10 %	71.50 %	72.30 %

Tabla 1.2: Evolución de la retención de primer año por sexo

Consistentemente las mujeres muestran una tasa de retención de primer año más alta que los hombres, observándose una diferencia de al menos 5 puntos porcentuales en los años registrados.

Establecimiento	**2014**	**2015**	**2016**	**2017**	**2018**
Municipal	68.8 %	69.4 %	70.6 %	72 %	73.1 %
P. Subvencionado	72.3 %	73.00 %	74.3 %	76.3 %	77 %
P. Pagado	78.9 %	77.7 %	77.9 %	79.9 %	80.4 %

Tabla 1.3: Evolución de la retención de primer año por dependencia del establecimiento

Se observa repetidamente, año tras año, que los estudiantes provenientes de un establecimiento particular pagado se ubican con mejores resultados a nivel de retención, seguidos por los particulares subvencionados y por último los de establecimientos municipales.

Sexo	2014	2015	2016	2017	2018
Mujeres	9.6	9.4	9.3	9.3	9.5
Hombres	10.5	10.2	10	10.1	10.3

Tabla 1.4: Evolución de la duración real de carreras por sexo

Al analizar la sobre-duración dentro de la educación superior agrupada por sexo, aparece una importante diferencia de casi 0.8 semestres en promedio año tras año.

Estos resultados indican la existencia de posibles correlaciones entre los fenómenos que el proyecto propone estudiar y los atributos característicos del estudiante, justificando así el uso de técnicas de minería de datos.

Con respecto a la experiencia científica, los estudios revelan la existencia de una multiplicidad de factores que contribuyen a explicar la aparición de estos comportamientos [7] [8] [9].

Por último, el proceso de deserción no es de total responsabilidad del estudiante, si no que también cumple un rol importante la institución académica [10], ya que la calidad de la institución es un factor relevante a considerar, sobre todo en los soportes que brindan al estudiante. También, destacan aspectos individuales, sociales e institucionales como desencadenante de la deserción universitaria, haciéndose notar la falta de una integración plena de los estudiantes, los problemas de calidad de la docencia y los sistema de apoyo institucional [11]

1.4. Definición del Problema

Las nuevas condiciones de la educación superior en Chile presentan un importante desafío, cómo es de esperar el incremento del acceso universitario permite el ingreso de un nuevo perfil de estudiantes, la expansión del sistema de educación superior se ha hecho con creciente énfasis en la integración de estudiantes pertenecientes a los quintiles de menores ingresos de la población [4]. Por lo tanto, el foco de la atención política ha cambiado de la cobertura hacia la efectividad del proceso formativo [12]. Hacerse cargo de este desafío hace necesario estudiar las dinámicas asociadas a los estudiantes que desertan o egresan oportunamente, buscando identificar estudiantes que se encuentren en riesgo y diseñando políticas de intervención.

El presente trabajo de título apoya la construcción de una herramienta que utilice análisis de datos para identificar estudiantes propensos a los fenómenos estudiados, por lo tanto, primeramente es necesario hacer una evaluación del funcionamiento de las distintas alternativas de manera local, es decir estudiando una casa de estudio en particular.

De este modo, el proyecto consiste en el estudio de la deserción temprana y el egreso oportuno de los estudiantes de ingeniería de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, utilizando para ello un conjunto de modelos de clasificación. Es necesario definir con mayor detalle los casos que son analizados.

Con respecto a la deserción temprana, se diferencian dos casos según el año en que el

estudiante hace abandono de sus estudios, la deserción correspondiente al primer año y al segundo año.

Respecto a la deserción del primer año solo se utilizarán atributos que son posibles de obtener al momento del ingreso del estudiante, esto quiere decir que se excluye cualquier información sobre los resultados académicos dentro de la universidad. Debido que es de interés aplicar los modelos al momento del ingreso.

Para la deserción correspondiente al segundo año, se utilizarán los mismos atributos del estudio de la deserción al primer año, solo que además se incluirá información correspondientes al rendimiento académico obtenido durante el primer año de estudio. De esta clasificación se excluyen los estudiantes que desertan al primer año.

Finalmente, se estudia el fenómeno de egreso oportuno, vale decir, que el estudiante logra completar su plan de estudio en el numero de semestres idóneo. Se estudia el egreso oportuno utilizando exclusivamente atributos conocidos a priori de su entrada a la universidad.

Para un correcto estudio de estos fenómenos se requiere contar con herramientas y modelos de minería de datos que sean pertinentes y confiables en la predicción de dichos fenómenos, por lo tanto, el éxito del estudio estará supeditado a la evaluación y comparación de los modelos, la selección de aquellos que ofrecen un mejor desempeño e identificación de atributos más relevantes.

Si bien, no todos los modelos llegan a resultados aceptables, ellos dependen en gran medida de los atributos que son analizados, siendo relevante la inclusión de variables de desempeño académico [13]. Además se señala la importancia de incluir variables socio-económicas [14], pero es menor la cantidad de estudios que analizan este tipo de variables, debido a que no es un atributo comúnmente registrado por las instituciones.

1.5. Objetivos

El proyecto a realizar se enmarca en el trabajo interuniversitario, mencionado anteriormente, el cual busca el desarrollo de una herramienta genérica para los gestores de permanencia de los establecimientos educacionales. Por lo tanto, por medio del presente trabajo se busca la aplicación y evaluación de distintos modelos de análisis de datos, con el fin de estudiar su rendimiento individual al explicar los fenómenos de deserción temprana y egreso oportuno en estudiantes de ingeniería de la Facultad de Ciencias Físicas y Matemáticas.

Objetivo General: " Evaluar el rendimiento de distintos modelos predictivos para el estudio de la deserción temprana y egreso oportuno para estudiantes de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile "

Objetivos Específicos:

- Categorizar estudiantes de la base de datos según historial académico
- Determinar métrica de desempeño adecuada para los fenómenos de deserción y egreso oportuno

- Determinar conjuntos de atributos relevantes para explicar los fenómenos de estudio
- Analizar los rendimientos obtenidos por parte de los modelos de clasificación
- Constrastar los resultados propios con los obtenidos en otras universidades

1.6. Metodología

El trabajo consiste en contrastar los modelos de clasificación, para lo cual se utilizará datos recopilados y proporcionados por la Universidad de Chile con el fin de encontrar patrones y/o relaciones entre los fenómenos en estudios y el conjunto de datos. Para esta tarea el área de análisis de datos ha desarrollado distintas metodologías, las que cuentan con un amplio grupo de herramientas que permiten extraer información de un conjunto de datos y transformarla en un conocimiento estructurado para su uso posterior.

S.E.M.M.A.

El desarrollo del proyecto se rige bajo la metodología SEMMA, utilizada para los procesos de minería de datos y propuesta por el SAS Institute, una de las compañías más importantes en el desarrollo de aplicaciones de software estadístico. SEMMA ofrece un proceso de minería de datos que consta de cinco pasos: Sample(muestreo), Explore (explorar), Modify (modificar), Model (modelar) y Assess (evaluar).

La etapa 'SAMPLE' corresponde al muestreo de los datos seleccionando el conjunto de datos apropiados para el modelado, este conjunto debe ser lo suficientemente grande para ser representativo y lo suficientemente pequeño para ser manejado por los algoritmos de manera eficiente, sin embargo gracias al desarrollo tecnológico esto ha sido menos relevante, ya que esta restricción se ha visto reducida gracias al mejoramiento de las herramientas computacionales.

La etapa 'EXPLORE' cubre la comprensión de los datos mediante el descubrimiento de relaciones anticipadas entre las variables, como también relaciones no anticipadas, además en esta fase se busca identificar anomalías dentro de los datos, para las tareas antes mencionada es de gran utilidad aplicar técnicas de visualización de datos.

En la etapa 'MODIFY' se emplean métodos para seleccionar, crear y transformar variables. Se limpian datos en caso que sea necesario como también se crean nuevas variables empleando la lógica del problema que se busca resolver. Al finalizar estas 3 primeras etapas de la metodología se busca tener como resultado un conjunto de datos limpio que puedan ser proporcionados a los algoritmos aprendizaje automático.

La fase de 'MODEL' tiene como enfoque la aplicación de técnicas de modelamiento que permitan obtener los resultados deseados, donde se busca combinaciones de reglas y patrones que predicen de manera confiable los fenómenos estudiados.

Por último, la etapa de 'ASSESS' corresponde a la evaluación de los modelos aplicados, señalando la confiabilidad y la utilidad que tienen los modelos desarrollados utilizando métricas de desempeño como criterio.

Aunque la metodología SEMMA contiene algunos de los elementos esenciales de cualquier proyecto de minería de datos, solo se refiere a las partes estadísticas del modelo y de la manipulación de datos. Esta metodología carece de algunas de las partes fundamentales de cualquier proyecto de sistemas de información, incluidas las fases de análisis, diseño e implementación, sin embargo , dada la naturaleza de la investigación, esta es apropiada, ya que solo se busca analizar y contrastar resultados.

1.7. Recurso de Datos

Para el desarrollo del trabajo de investigación se cuenta con una base de datos de 5,945,906 de filas que registran los atributos de a 172,935 estudiantes de la Universidad de Chile. Dentro de este conjunto de variables se cuenta con 31 atributos, siendo posible identificar cuatro grupos de atributos, variables socio-demográficas, variables socio-económicas, variables académicas y variables de postulación. Esta cuenta con datos recopilados hasta el año 2017.

1.8. Resultados Esperados

Los resultados esperados para este trabajo consiste en la elaboración de un registro detallado sobre los modelos predictivos que fueron aplicados para los distintos casos de estudios. Estos deben incluir métricas de desempeño, en conjunto con los atributos que permiten alcanzar dicho nivel de rendimiento. Además de identificar qué modelo se muestra con mejores resultados para explicar los fenómenos entre la cartera de modelos.

1.9. Alcances y Limitaciones

El proyecto se realiza sobre estudiantes pertenecientes a la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, por lo tanto las conclusiones y resultados obtenidos pueden ser extrapolados a instituciones con características similares, ya que es un establecimiento singular dentro del sistema.

Por otra parte, el campo de la investigación sobre este tipo de fenómenos se presenta algo confuso, principalmente porque no existe una definición bien específica entre los investigadores sobre qué es lo que se debe denominar deserción estudiantil. Esto ha traído consigo la aparición de estudios que afirman que la probabilidad de que un estudiante interrumpa sus estudios disminuye en la medida que el estudiante presenta mejores cualidades académicas, mientras otras investigaciones sostienen que alumnos más brillantes tienden a abandonar sus estudios. Estas dos ideas opuestas son solo un reflejo de lo importante que es una correcta definición del fenómeno para la investigación [15].

El estudio comprende la aplicación de modelos predictivo sobre la base de datos de estudiantes de la Universidad de Chile, con el objetivo de estudiar los casos de interés dentro del paso del estudiante por la educación superior, sin embargo, estas dificultades no necesariamente corresponden a una mala respuesta de cara a las exigencias académicas, ya que se reconoce la existencia de estudiantes que no tienen suficiente interés, algunos no están completamente comprometidos con la finalización de sus estudios, y por lo tanto la deserción

o retraso en esta clase de estudiante corresponde más a una falta de interés que a dificultades académicas en desarrollo de su carrera universitaria [16]. Por lo tanto al no poder tener registro cualitativo ni cuantitativo de este factor es imposible poder identificar a que clase corresponde el estudiante.

Además, como lo muestra la literatura, los factores económicos pueden representar atributos significativos al momento de explicar esta clase de fenómenos, sin embargo nuevamente se presentan limitantes por la disposición de atributos con los que cuenta la base de datos.

Capítulo 2

Revisión Bibliográfica

A continuación, se presenta una serie de investigaciones ya realizadas que abordan los problemas de los que se hace cargo el presente trabajo, la deserción universitaria y el egreso en el tiempo apropiado. Si bien, es posible reconocer una similitud entre estos fenómenos, estos son tratados generalmente de manera independiente.

2.0.1. Deserción o Retención Universitaria

Reyes, Escobar, Duarte y Ramirez (2007) utilizan un modelo de regresión logística para predecir el éxito de los estudiantes en el primer semestre de la carrera, utilizando variables explicativas como la expectativa del alumno sobre su rendimiento, la importancia hacia los resultados, la valoración del esfuerzo propio, además de variables como las notas de enseñanza secundaria y los puntajes obtenidos en la prueba de selección universitaria. Encontrando más significante las variables de notas escolares y los puntajes de selección en los módulos de matemáticas y ciencias. Por otra parte, Esteban (2017) encuentra que el rendimiento académico en la universidad es la variable más influyente en la permanencia del estudiante dentro de la institución.

Himmel (2002) detecta que un importante número de estudiantes que abandonan sus estudios se localizan dentro de los primeros años de estudios, siendo menor la proporción de estudiantes que desertan a medida que avanzan los años dentro del sistema de educación superior. De manera complementaria, Stinebrickner (2014) encuentra que el 45 % de los abandonos en los primeros dos años de universidad son atribuidos al rendimiento académico, aunque este efecto se va reduciendo en los años académicos posteriores.

De igual manera, el cese de los estudios universitarios puede ser explicado por múltiples factores, no existiendo un único factor causal para que el estudiante tome la decisión de desertar. Estudios refuerzan la importancia de la influencia social en los estudiantes de pregrado, en particular el efecto que tienen sus mismos pares dentro del aula, como también la influencia que generan sus padres, siendo estos últimos menos relevantes en la medida que el estudiante avanza en el sistema educacional [17]. Los factores sociales influyen en la decisión de ingresar a la educación superior, momento donde deben eligir la carrera universitaria e institución, así no existiendo igualdad de oportunidades entre los estudiantes, viéndose limi-

tados a restricciones como aranceles, becas, su ubicación, entre otras. Estas son variables que limitan las oportunidades de elección.

De la misma forma, existe preferencia por parte de los estudiantes a ingresar a instituciones educativas en las que encontrarán gente similar a ellos, en busca de un sentido de pertenencia [18]. Esto da cuenta la importancia que tiene el entorno en donde el alumno realiza sus estudios, tomando un rol activo el factor institucional. Spady (1970) y Tinto (1975) ya mencionan la integración del estudiante como un aspecto relevante para explicar la deserción, sugiriendo que estos hacen abandono cuando ya no existe una pertenencia al sistema académico social.

Gonzalez y Uribe (2005) reafirman los factores institucionales y sus aspectos sociales. Señalando que las instituciones no han logrado adaptarse a los nuevos tipos de estudiantes que empiezan a ingresar a la educación superior, además indica factores motivacionales y factores económicos como gatillantes para que el estudiante tome la decisión de desertar.

Donoso y Schiefelbein (2007) plantean que los problemas de deserción provienen de las injusticias sociales, donde los alumnos menos privilegiados se encuentran en desventaja con respecto a sus pares que provienen de familias más acomodadas. De esta forma, se considera a un mejor estudiante, no por sus habilidades de tipo intelectual, si no que estas se encuentran asociadas con un nivel de capital social, cultural, económico y educacional previo. Siendo el nivel socio-económico y el capital cultural de las familias los principales factores que explican las diferencias de rendimiento.

2.0.2. Tiempo Hasta el Egreso del Estudiante

Diblasi (2005) estudia cuales son las probabilidades del estudiante de finalizar y egresar de su plan de estudios, a través del estudio de una población de estudiantes que ingresan entre el 1997 y el 2003 a la Universidad Nacional de Cuyo, Argentina. Encontrando que la probabilidad de éxito se encuentra significativamente influenciada por aspectos como la edad, sexo, tipo de establecimiento escolar de procedencia, nivel educacional de los padres y carrera escogida.

Yue y Fu (2017) utilizan modelos de riesgo en tiempos discretos para estudiar el fenómeno, utilizando atributos más allá de las características preuniversitarias y revelando que el rendimiento académico es el factor más importante seguido por la decisión sobre las especialidades que toma el alumno.

Lassibille y Navarro Gómez (2011) muestran que el tiempo de finalización es extremadamente sensible a las habilidades de los estudiantes y, en mucho menor grado, al contexto socioeconómico, la motivación al ingresar al programa y el género. Las habilidades académicas previas a la matricula son una determinante significativa en el tiempo hasta la graduación, independientemente del programa de estudio. Por otra parte, la edad en el momento del ingreso a la institución se relaciona positivamente con un progreso lento hacia el grado. Mientras, Lovenheim y Turner (2012) muestran que las tardanzas son más comunes en estudiantes que no se encuentran preparados académicamente.

Swail (2003) examina la relación entre las prácticas institucionales y las necesidades aca-

démicas y sociales de los estudiantes, encontrando que las instituciones que trabajan de manera proactiva para ayudar a los estudiantes a tener éxito y satisfacer sus necesidades pueden acortar el tiempo para graduarse.

Otras investigaciones destacan la relación de el aumento en los tiempos de egreso a factores más complejos, incluyendo la reducción de recursos institucionales disponibles para los estudiantes, aumento de los costos universitarios [19], necesidad de planes de nivelación [20], una mayor inscripción a los cursos requeridos, a cursos que no corresponden al grado y una dificultad de obtener los cursos requeridos; necesidades de empleo; entre otros[21].

El tiempo que demora el estudiante en graduarse son preocupaciones primordiales en la educación superior, sin embargo la comprensión aún es limitada con respecto a los factores que afectan el tiempo hasta el grado [22]. Por ejemplo, Knight y Arnold (2000) atribuyen la culpa del tiempo prolongado a las instituciones, mientras que otros estudios lo atribuyen a las decisiones individuales del estudiante. La falta de comprensión se debe principalmente a dos barreras: disponibilidad de datos y metodología [23], es difícil hacer un seguimiento del progreso de los estudiantes a lo largo del tiempo, además no existe una metodología acordada para modelar los efectos del progreso académico de los estudiantes en el tiempo hasta el grado.

Capítulo 3

Marco Conceptual

3.1. Aprendizaje Automático

El aprendizaje automático corresponde una colección de algoritmos y técnicas utilizadas para crear sistemas computacionales que obtienen predicciones a partir de datos recopilados. Los algoritmos de aprendizaje automático crean un modelo basado en datos de muestra, conocidos como 'datos de entrenamiento', que permiten realizar predicciones sobre nuevos conjuntos más amplios al identificar patrones o tendencias.

Los algoritmos utilizados dentro del aprendizaje automático pueden ser clasificados en tres grupos, aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. Dentro del presente proyecto se desarrollan modelos perteneciente al primer y último grupo.

La construcción de estos modelos poseen tres etapas:

- Captación, se debe proporcionar al modelo los datos de entrada y el valor de salida asociado, para que este pueda reconocer patrones dentro de un conjunto datos (datos de entrenamiento), permitiendo la construcción de una función estadística conocida comúnmente como modelo.
- Validación, en esta siguiente fase, se evalúa el rendimiento de dicha función utilizando, generalmente, el conjunto restante de datos no utilizados en la fase anterior. La evaluación se realiza por medio alguna métrica de rendimiento definida previamente.
- Predicción, por último se aplica el modelo entrenado a nuevos conjuntos de datos.

3.1.1. Aprendizaje Supervisado

Para el aprendizaje supervisado se proporciona un conjunto de datos de entrada y de salida. Con el objetivo de que el modelo 'aprenda' de los patrones entregados, construyendo así, un conjunto general de reglas que permite entregar un valor de salida a nuevos valores de entrada.

Existen dos tipos de algoritmos de aprendizaje supervisado de uso común, uno de ellos es de regresión, donde los resultados a predecir corresponden a un valor numérico, mientras que

los algoritmos de clasificación predicen pertenencia a alguna clase.

Logistic Regression

La regresión logística es un algoritmo de aprendizaje automático que se utiliza para los problemas de clasificación. Esta es apropiada de realizar cuando la variable dependiente es binaria.

Una regresión logística es un modelo de regresión lineal, pero utilizando una función de costo más compleja, esta función de costo puede definirse como la 'función sigmoidea' en lugar de una función lineal. La hipótesis de la regresión logística tiende a limitar la función de costo entre 0 y 1. Así, este es un algoritmo de análisis predictivo que se basa en el concepto de probabilidad de pertenencia de clase.

Decision Tree

Un árbol de decisión es una representación simple para clasificar datos. Es un aprendizaje automático supervisado donde los datos se dividen continuamente de acuerdo a un determinado parámetro.

Un árbol de decisión consta de:

- Nodos: Evalúa el valor de un determinado atributo
- Rama: Flujo dependiente del resultado de una prueba que se conecta al siguiente nodo u hoja
- Hoja: Nodos terminales que predicen el resultado, representando la etiqueta de clase

Para estos algoritmos, la bondad de división se cuantifica por una medida de impureza, por lo tanto se dice que la división es pura si todos los datos seleccionados de una rama pertenecen a la misma clase.

Este es probablemente el modelo más popular debido a su facilidad de interpretación, permitiendo clasificar un ejemplo concreto comenzando en su raíz y siguiendo un camino determinado por las ramas según la respuesta a las preguntas que se encuentran en los nodos, donde finalmente debe terminar en una hoja del algoritmo que proporciona el resultado.

Naive Bayes

Un clasificador Naive Bayes es un modelo probabilístico de aprendizaje automático que se utiliza para la tarea de clasificación.

Usando el teorema de Bayes, podemos encontrar la probabilidad de que ocurra el suceso A, dado que el suceso B ha ocurrido. Aquí, B es el conjunto de atributos (evidencia) y A es la la clase que representa (hipótesis). La suposición hecha aquí es que los predictores son independientes, es decir, la presencia de una característica particular no afecta a la otra.

Teoremas de Bayes:

$$P(C \mid A) = \frac{P(A \mid C)\, P(C)}{P(A)}$$

Al ser B el conjunto de atributos, este se puede expresar como:

$$A_1, A_2, A_3, ..., A_n$$

, Por lo tanto la expresión se puede escribir como:

$$P(C = c \mid A_1 = a_1, A_2 = a_2, ..., A_n = a_n) = \frac{P(A_1 = a_1, A_2 = a_2, ..., A_n = a_n \mid C = c)\, P(C = c)}{P(A_1 = a_1, A_2 = a_2, ..., A_n = a_n)}$$

Estimar el primer término del denominador puede resultar bastante complejo, por esto de manera ingénua, supone que los atributos son independientes, y por lo tanto, utilizando propiedades de probabilidades la ecuación se puede escribir de la siguiente forma:

$$P(C = c \mid A_1 = a_1, A_2 = a_2, ..., A_n = a_n) = \frac{\prod_{\mathrm{i}=1}^{n} P(A_\mathrm{i} = a_\mathrm{i} \mid C = c)\, P(C = c)}{P(A_1 = a_1, A_2 = a_2, ..., A_n = a_n)}$$

No se requiere calcular el denominador porque este valor se mantiene fijo para los distintos valores que toman las clases, luego el numerador suministra la información necesaria para determinar que clase tiene mayor probabilidad. De modo que la clasificación consiste en determinar la clase que maximiza la expresión del numerador, más fácil de calcular por la suposición de independencia.

K Nearest Neighbors

K vecinos más cercanos es un algoritmo simple de clasificación supervisada, que almacena todos los casos disponibles y clasifica los nuevos en función de una medida de similitud (por ejemplo, funciones de distancia).

El algoritmo no hace ninguna suposición sobre la distribución de las variables predictivas. En la fase de entrenamiento se almacenan los vectores característicos y las etiquetas de pertenencia de clase, en la fase de evaluación un nuevo registro se representa como un vector en el espacio y se seleccionan los K ejemplos más cercanos, donde el nuevo registro es clasificado como la clase que más recurrente entre sus vecinos.

Este método supone que los vecinos mas cercanos entregan una buena clasificación de la etiqueta, sin embargo dicha suposición presenta un problema cuando se considera una cantidad dominante de atributos irrelevantes, haciendo que los atributos relevantes pierdan influencia frente a estos.

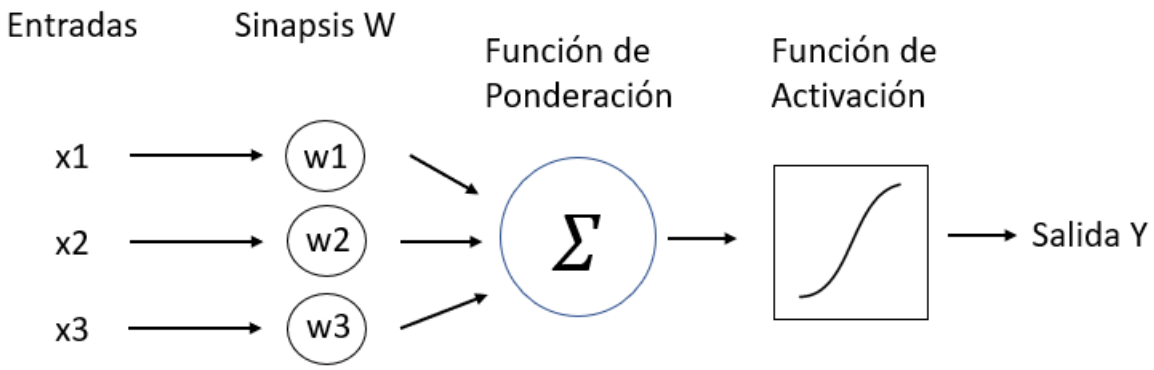
La elección del valor óptimo para K se realiza mejor inspeccionando primero los datos. En general, un valor K grande es más preciso, ya que reduce el ruido general, pero no hay garantía de aquello. Históricamente, la K óptima para la mayoría de los conjuntos de datos ha estado entre 3-10.

Neural Net

Una red neuronal consiste en distintas neuronas dispuestas en capas, las que convierten un vector de entrada en un vector de salida. Estas redes son conformadas de manera secuencial, es decir una neurona alimenta a la siguiente que se ubica en la próxima capa.

Poseen un funcionamiento análogo al funcionamiento de una neurona biológica, la neurona i-esima recibe señales de las neuronas adyacentes de la capa anterior, además estas señales son ponderadas por distintos pesos asignados por el modelo. La activación de la neurona se da siempre y cuando la suma de las señales ponderadas superen un valor, conocido como umbral de activación, en cuyo caso esta entrega un valor de salida. Este funcionamiento se aprecia de manera gráfica en la siguiente imagen:

Figura 3.1: Representación ilustrativa del funcionamiento de una neurona artificial

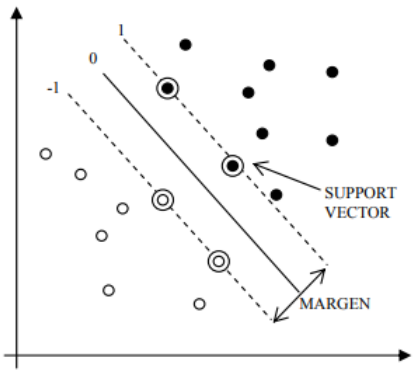


Super Vector Machine

El modelo de Super Vector Machine se basa en encontrar un hiperplano separador de clases, de tal forma que se maximice la distancia entre dos hiperplanos construidos de manera paralela y a cada lado del hiperplano separador. De esta forma se definen las regiones correspondiente a cada clase de estudio. Estos hiperplanos formulados son denominados vectores de soporte.

De manera gráfica se puede observar el siguiente gráfico, el que corresponde un problema de clasificación binaria.

Figura 3.2: Ejemplo de clasificación binaria



El objetivo del método es encontrar el hiperplano de separación que maximiza el margen, la distancia entre los hiperplanos paralelos, quedando así ambas regiones justamente definida según los datos estudiados.

3.1.2. Métodos de ensamblaje

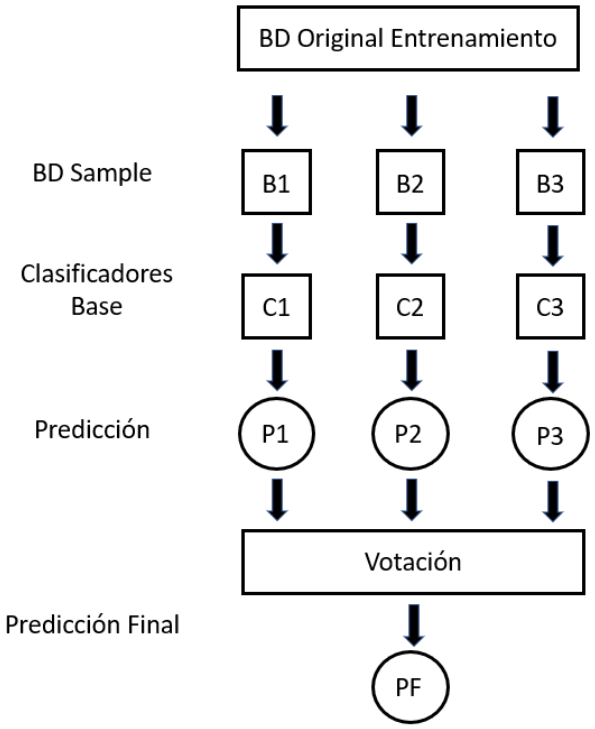
Los métodos de ensamblaje permiten mejorar el rendimiento predictivo mediante la combinación de diversos modelos. Estos métodos son meta-algoritmos que combinan los poderes predictivos de otros modelos para disminuir la varianza o sesgo de sus predicciones. Esto concepto se basa en la idea de tomar una decisión final combinando las distintas decisiones individuales.

Sin embargo, es importante enfatizar que no hay garantía de que la combinación de clasificadores múltiples funcione mejor que él clasificador individual. Por lo tanto, no se puede garantizar una mejora en el rendimiento promedio de los modelos utilizados, excepto en ciertos casos especiales [24].

Bootstrap Aggregation (Bagging)

Este método contempla el uso de diferentes subconjuntos extraídos al azar (con reemplazo) sobre el conjunto de datos de entrenamiento. Cada subconjunto se usa para entrenar un clasificador diferente del mismo tipo. Los clasificadores individuales se combinan mediante la suma de votos en sus decisiones de clasificación, por lo tanto la clase escogida por el clasificador ensamblado corresponde la clase con mayor número de repeticiones entre los clasificadores individuales.

Figura 3.3: Ejemplo de método bagging

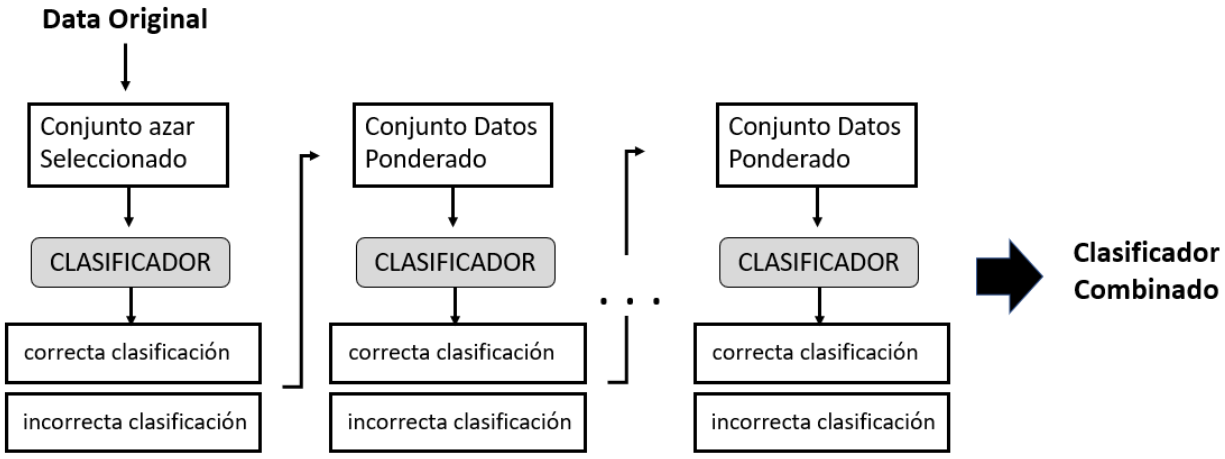


AdaBoost Method (Boosting)

Ada-boost o Adaptive Boosting combina múltiples clasificadores para aumentar el rendimiento de la clasificación, este construye un clasificador fuerte al combinar varios clasificadores debiles. Los clasificadores débiles son clasificadores que funcionan levemente mejor que un clasificador aleatorio, pero aún tiene un desempeño pobre en la asignación de clases.

El algoritmo Boosting intenta construir un clasificador fuerte a partir de los errores de varios modelos más débiles. Así, comienza por crear un modelo a partir de los datos de entrenamiento, luego crea un segundo modelo a partir del anterior tratando de reducir los errores del modelo. Los modelos se agregan de forma secuencial, cada uno corrigiendo a su predecesor, hasta que los datos de entrenamiento se predigan perfectamente o se haya agregado el número máximo de modelos.

Figura 3.4: Ejemplo de método boosting

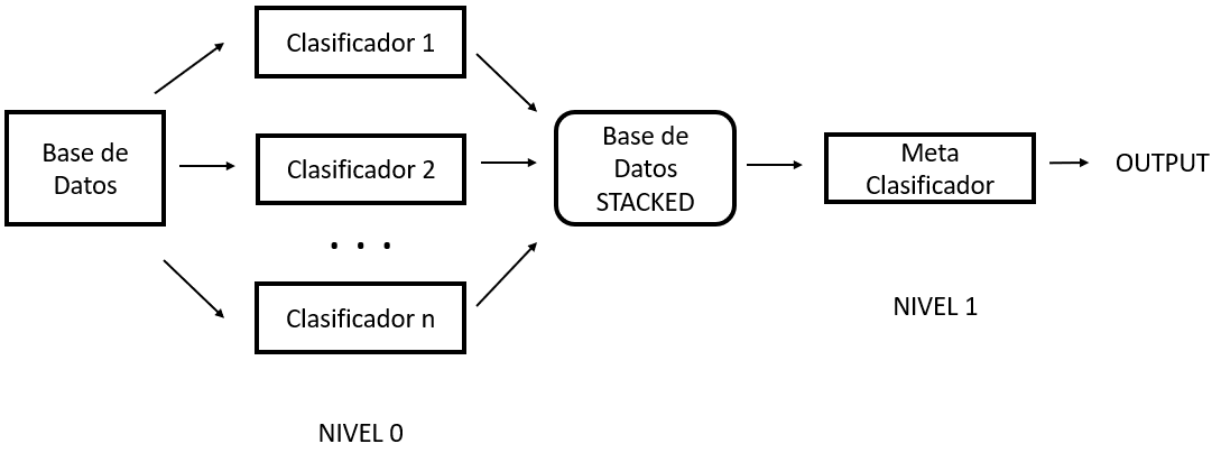


Stacked Generalization

Stacked Generalization corresponde a una técnica de ensamblado que usa un nuevo modelo para aprender cómo combinar mejor las predicciones de dos o más modelos. Se crean clasificadores de Nivel 0 , cuyos resultados son utilizados para entrenar un clasificador de Nivel 1, un meta clasificador.

La noción corresponde a saber si los modelos implementados en el nivel 0 han aprendido correctamente de los datos de entrenamiento, si un clasificador particular aprendió incorrectamente una determinada región del espacio, entonces el clasificador de nivel 1 puede aprender este comportamiento y, junto con los comportamientos aprendidos por los otros clasificadores, puede corregir ese entrenamiento inadecuado.

Figura 3.5: Ejemplo de método stacked generalization



3.2. Bases de Datos Desbalanceadas

La mayoría de los problemas del mundo real muestran un cierto nivel de desequilibrio de clases, donde las clases con mayor interés para las investigaciones se encuentran en menor frecuencia en las bases de datos.

Cuando se cuenta con un conjunto de datos desbalanceados, es más difícil para un modelo aprender las características de los ejemplos pertenecientes a la clase minoritaria, debido a la escasez de ejemplos en las bases de datos, haciendo que la clasificación se encuentre sesgada.

Por lo tanto, es imprescindible realizar ajustes al modelo antes de ser entrenado, además de seleccionar una métrica apropiada de rendimiento, de no ser llevadas estas tareas a cabo se predispone a conseguir malos resultados o conclusiones erradas. Un ligero desequilibrio a menudo no es una preocupación, y el problema se puede tratar como un problema de modelado predictivo de manera normal, sin embargo un desequilibrio severo de las clases puede ser difícil de modelar y puede requerir el uso de técnicas especializadas.

3.3. Evaluadores de Desempeño

Para tener criterio sobre la calidad de los modelos aplicados es necesario la definición de evaluadores de desempeños. En particular, cuando se utilizan modelos de clasificación, es útil recurrir a una matriz de confusión.

Una matriz de confusión corresponde a la representación matricial de los verdaderos positivos (TP, del inglés True Positive) y negativos (TN, del inglés True Negative), correspondientes a los aciertos del clasificador, y los falsos positivos (FP, del inglés False Positive) y negativos (FN, del inglés False Negative), que corresponden a errores de clasificación.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	True Positives	False Negatives
	Negativos	False Positives	True Negatives

Tabla 3.1: Matriz de confusión

- TP: número de clasificaciones positivas del modelo que coinciden con los valores positivos del set evaluado (clasificaciones positivas correctas).
- TN: número de clasificaciones negativas del modelo que coinciden con los valores negativos del set evaluado (clasificaciones negativas correctas).
- FP: número de clasificaciones positivas del modelo que no coinciden con los valores positivos del set evaluado (clasificaciones positivas incorrectas).
- FN: número de clasificaciones negativas del modelo que no coinciden con los valores negativos del set evaluado (clasificaciones negativas incorrectas).

Accuracy : Medida de clasificación que evalúa el número de clasificaciones correctas sobre el total de clasificaciones realizadas. Esta métrica toma valores entre 0.0 y 1.0, donde un valor cercano a 1.0 significa que el modelo predice correctamente, tanto clasificaciones positivas como negativas.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Estas métricas permiten tener una forma simple de describir el rendimiento del clasificador. Sin embargo, el valor obtenido puede ser engañoso en algunas situaciones, en particular en estudios sobre conjuntos de datos desbalanceados.

A modo de ejemplificar se puede suponer la situación donde la clase mayoritaria representa el 95 % de las muestras, mientras que la clase minoritaria representa solo el 5 %. Un clasificador que prediga toda las instancias hacia la clase mayoritaria proporciona una métrica del 0.95. Sin embargo, este valor que aparentemente corresponde a un buen rendimiento, corresponde al hecho de no lograr identificar ninguna muestra de la clase minoritaria. Otras métricas de desempeño que se adoptan con frecuencia en problemas de aprendizaje con datos desbalanceados corresponden a precision, recall y f1-score.

Precision : Esta medida de clasificación cuantifica el porcentaje de predicciones positivas realizadas correctamente. En un problema de clasificación binaria, la precisión se calcula como el número de verdaderos positivos dividido por el número total de verdaderos positivos y falsos positivos.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Recall : Métrica que representa la razón entre en número identificados de clases positivas sobre el total de casos de clase positiva. Este valor se calcula como el número de verdaderos positivos divididos por el número total de verdaderos positivos y los falsos negativos.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

F1 - Score : Medida de clasificación que evalúa el porcentaje de clasificaciones correctas, enfocada solo en la clase positiva (el significado de la clase positiva puede variar, dependiendo de la variable objetivo). Esta métrica, normalmente se utiliza cuando la proporción de una de las dos clases es significativamente mayor a otra. El F1-score toma valores entre 0.0 y 1.0, significando un valor alto que el modelo es capaz de predecir correctamente la clase positiva.

El f1-score se usa para medir el rendimiento de un clasificador equilibrando el uso de precision y recall. De esta forma, la métrica puede proporcionar una medida más realista, ya que tiene en cuenta tanto los falsos positivos como los falsos negativos. Se obtiene del cálculo de la media armónica entre precision y recall.

$$F - Score = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.4}$$

3.4. Cross Validation

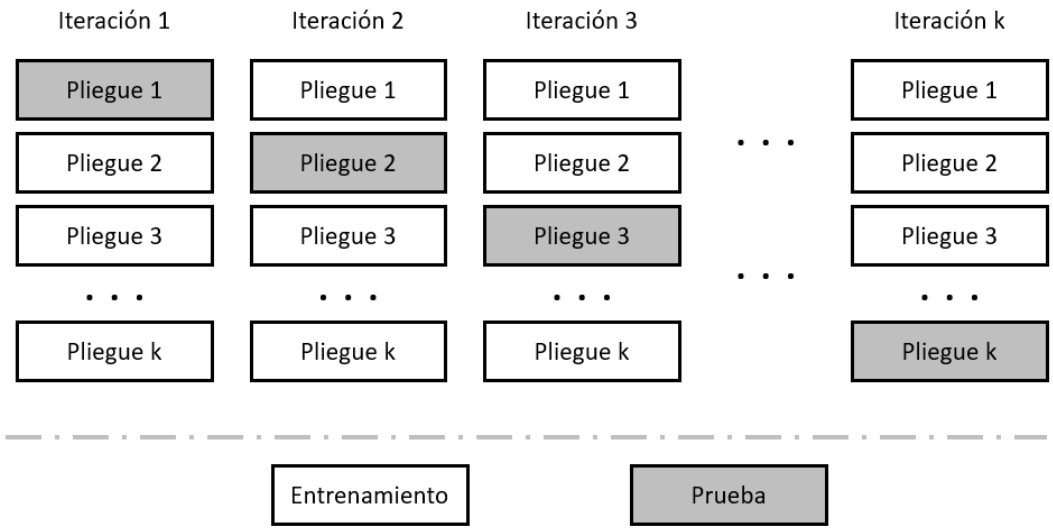
Para el ajuste de un modelo se requiere utilizar un conjunto de datos de entrenamiento para luego ser evaluado sobre el conjunto de datos prueba. El sesgo y la varianza siempre existen en los algoritmos supervisado de aprendizaje automático, debido a que el conjunto de datos de entrenamiento generalmente no puede incluir todos los ejemplos posibles. La razón principal para utilizar validación cruzada corresponde a intercambiar el sesgo por una varianza baja, y así disminuir el impacto de la selección de datos de entrenamiento.

La técnica consiste en realizar diferentes particiones que conforman los grupos de entrenamiento y de prueba, para después calcular la media de los rendimientos obtenidos.

Para el trabajo de investigación se utiliza el caso especial llamado Leave-one-out, por lo tanto el conjunto de datos se divide en k subconjuntos, y el método se repite k veces. En cada iteración, uno de los k subconjuntos se usa como conjunto de prueba y los otros subconjuntos k-1 se unen para formar un conjunto de entrenamiento. Luego, se calcula la métrica de rendimiento en cada una de las k pruebas y se promedian.

La ventaja de este método es que se ve disminuido el efecto que provoca la división de los datos, utilizando todos los datos disponibles para probar el rendimiento del modelo. Sin embargo, la desventaja de este método radica en que el algoritmo de entrenamiento se debe volver a ejecutar k veces, lo que significa que se necesitan k veces más cálculos para realizar la evaluación.

Figura 3.6: Representación gráfica de partición por método de validación cruzada



Capítulo 4

Metodología

4.1. Estructuración de Datos

Como se mencionó anteriormente, para realizar la investigación se cuenta con un data set de 5.945.906 registros, donde se identifican a 172.934 estudiantes únicos pertenecientes a las distintas facultades de la Universidad de Chile. Sin embargo, para la investigación solo se contempla a los estudiantes de la Facultad de Ciencias Físicas y Matemáticas (FCFM), teniendo registro de 10.413 estudiantes con información hasta el año 2017. Estos pueden estar cursando una de las distintas carreras que se imparten dentro de la facultad, las cuales poseen una duración de 12 semestres, a excepción de ingeniería civil en matemáticas y computación con 11 semestres y las carreras de licenciatura en ciencias (Astronomía, Geofísica y Física) con duración de 10 semestres.

Los estudiantes al ingresar a la FCFM deben cursar de manera obligatoria un programa común para todas las ingenierías y licenciaturas, este tiene una duración de dos años y permite construir un sólido desarrollo en las ciencias básicas.

Carrera Universitaria	Duración de Semestres
GEOLOGÍA	12
INGENIERÍA CIVIL	12
INGENIERÍA CIVIL DE MINAS	12
INGENIERÍA CIVIL ELÉCTRICA	12
INGENIERÍA CIVIL EN BIOTECNOLOGÍA	12
INGENIERÍA CIVIL EN COMPUTACIÓN	11
INGENIERÍA CIVIL INDUSTRIAL	12
INGENIERÍA CIVIL MATEMÁTICA	11
INGENIERÍA CIVIL MECÁNICA	12
INGENIERÍA CIVIL QUÍMICA	12
ASTRONOMÍA (LICENCIATURA)	10
GEOFÍSICA (LICENCIATURA)	10
FÍSICA (LICENCIATURA)	10

Tabla 4.1: Carreras impartidas en la Facultad de Ciencias Físicas y Matemáticas

La base de datos cuenta con 31 atributos, estos pueden ser características propias del estudiante como también atributos correspondientes al resultado de un determinado periodo. La base está construida por medio del cruce de información de diferentes bases de datos por medio de consultas que no responden exactamente a los requerimientos del proyecto, esto provoca que los atributos no sean únicos y que exista duplicidad de información. Por lo tanto, se requiere realizar una re-estructuración de datos para poder extraer la información útil para la investigación.

Se construyen dos bases de datos, ambas registrando atributos para los 10.413 estudiantes únicos. Una de estas queda formada por atributos característicos del estudiante, mientras que la otra se construye a partir del historial académico, esta es necesaria para identificar la clase de pertenencia en los distintos casos de estudio, deserción temprana o egreso oportuno.

Atributos Demográficos

- Rut
- Edad de ingreso
- Género
- Comuna
- Provincia
- Región

Atributos del Establecimiento

- Dependencia educacional
- Tipo de educación

Atributos del Proceso de Ingreso

- Año de ingreso
- Preferencia
- PSU
- PAA
- Beca excelencia académica (BEA)
- Pertenencia a programa equidad de género (PEG)
- Programa para deportistas destacados
- Cupo estudiante extranjero
- Prog. acceso efectivo a la ed. superior (PACE)
- Ingreso prioritario equidad educativa (SIPEE)

Atributos de Rendimiento Académico Pre Ingreso

- Promedio enseñanza media
- Puntaje NEM
- Puntaje ranking
- Puntaje PSU matemáticas
- Puntaje PSU lenguaje
- Puntaje PSU historia
- Puntaje PSU ciencias
- Puntaje PSU ciencias biología
- Puntaje PSU ciencias física
- Puntaje PSU ciencias química
- Puntaje ponderado

Atributos de Historial Académico

- Ramos reprobados en el primer año
- Situación académica

4.2. Identificación de Casos de Estudios

La investigación se centra en estudiar el comportamiento de 3 casos de estudios, para esto se recurre al historial académico del estudiante, construido a partir la base de datos original.

1. Para identificar a los estudiantes que desertan después del primer año de estudio, se considera a quienes no posean información académica correspondiente al año siguiente del ingreso, es decir, su segundo año académico, por lo tanto son considerados, tanto desertores permanentes como desertores temporales. Los estudiantes que ingresan el año 2017 no son etiquetados, ya que no se puede conocer de su comportamiento futuro al no existir la información.
2. Para identificar a los estudiantes que desertan después del segundo año, se considera quienes cursaron apropiadamente sus primero dos años (excluye desertores de primer

año), pero que no presentan registro al año siguiente, de igual forma, estos pueden ser tanto desertores permanentes como temporales. Los estudiantes que ingresan el año 2016 y 2017 no son etiquetados, ya que no se puede conocer de su comportamiento futuro al no existir la información.

3. Para identificar a los estudiantes que egresaron oportunamente, primero se consideran como candidatos quienes muestran una trayectoria continua, es decir sin haber desertado en algún año, e ingresan antes del 2012, debido a la disponibilidad de información. Así son considerados como estudiantes con egreso oportuno quienes finalizan sus estudios al sexto año. Para este caso de estudio se excluye a los estudiantes que cursan licenciaturas, debido a una diferencia en la duración de las carreras con el común de la facultad.

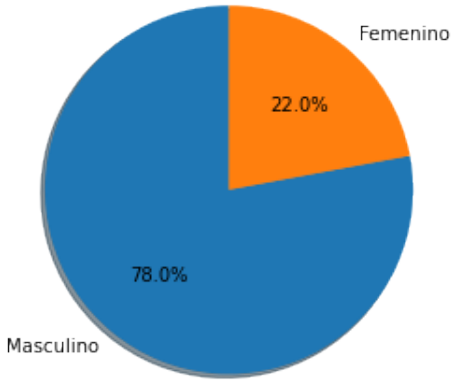
4.3. Análisis Exploratorio

A continuación se presenta un primer análisis de carácter exploratorio sobre las variables que son recopiladas, con el fin de visualizar el comportamiento de estas y asegurando su utilidad para la investigación.

GENERO:

Se presentan la distribución porcentual entre los estudiantes masculinos y femeninos, donde solo 1 de cada 5 estudiantes es del género femenino. Sin embargo, pese a lo bajo de este valor, esto se debe a una tendencia general que tienen las carreras de ingeniería [28].

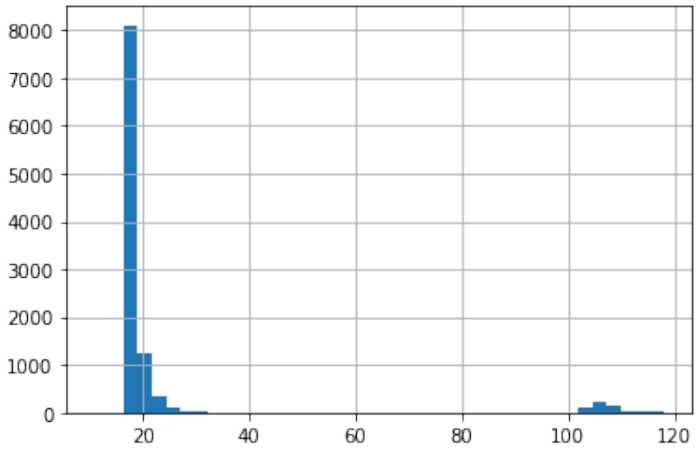
Figura 4.1: Distribución de género en el total de estudiantes



EDAD DE INGRESO:

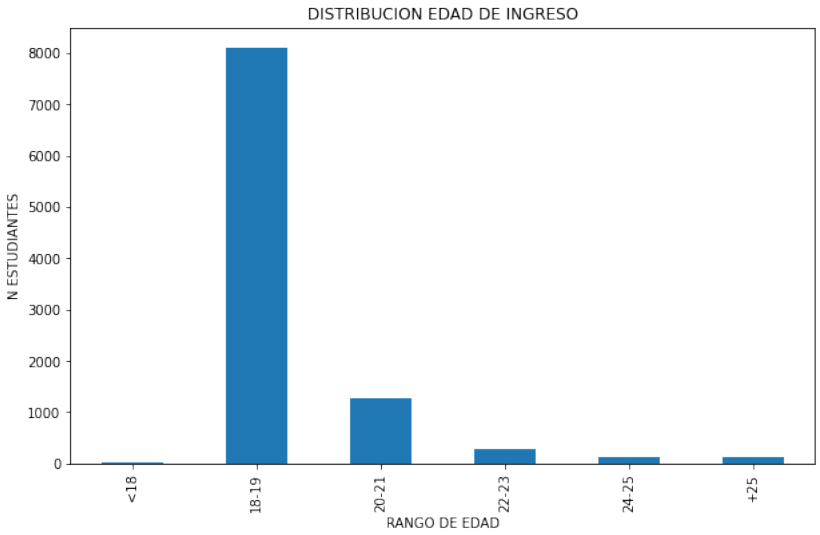
Existen estudiantes con valores atípicos en la edad de ingreso, teniendo registro de estudiantes que ingresan a la universidad posterior a cumplir sus 100 años, por lo tanto, se opta por eliminar estos valores del atributo.

Figura 4.2: Distribución edad de ingreso



Se realiza una visualización por rango de edad, se observa que más de 8,000 estudiantes de los 10,413 se encuentran dentro del rango de 18 y 19 años. Este atributo presenta un promedio de 19.19 años con una desviación estándar de 1.64 años.

Figura 4.3: Distribución de edad de ingreso por rango



REGIÓN:

El atributo región corresponde a la región del domicilio registrado por el estudiante, se observa una alta concentración de estudiantes pertenecientes a la Región Metropolitana. Si bien, el atributo no coincide necesariamente con la región de procedencia del estudiante, si coincide en cierto grado con dicha información, por lo que al no contar con la información actualizada se permitirá su uso.

Por otra parte, se tienen 51 estudiantes sin informar región.

REGIÓN	ALUMNOS
Metropolitana de Santiago	9625
Del Libertador B. O'Higgins	238
De Valparaíso	123
Del Maule	88
De Coquimbo	54
No Informado	51
De La Araucanía	43
De Los Lagos	43
Del Ñuble	33
De Antofagasta	32
Del Bíobío	29
De Los Ríos	19
De Magallanes y de La Antártica Chilen	13
De Atacama	9
De Tarapacá	7
De Arica y Parinacota	5
De Aisén del Gral. C. Ibáñez del Campo	1

Tabla 4.2: Número de estudiantes por región de domicilio

PROVINCIA:

La base de datos cuenta con 47 provincias distintas, además se tienen 51 estudiantes con provincia no informada. Se muestra una gran concentración en estudiantes pertenecientes a la provincia de Santiago, sin embargo, al igual que en el caso anterior este atributo no corresponde necesariamente a la provincia de procedencia, si no a la de domicilio.

PROVINCIA	N. ALUMNOS
Santiago	8711
Cordillera	353
Maipo	247
Cachapoal	212
Talagante	150
Chacabuco	109
Melipilla	55
No Informado	51
Elqui	47
Valparaíso	40

PROVINCIA	N. ALUMNOS
Cautín	39
Curico	36
Talca	34
Diguillín	26
Llanquihue	24
Los Andes	22
Concepción	21
Colchagua	21
Antofagasta	20
San Antonio	19

PROVINCIA	N. ALUMNOS
Valdivia	18
Linares	17
San Felipe	14
Marga Marga	13
Magallanes	13
El Loa	11
Quillota	11
Osorno	9
Bío- Bío	8
Chiloe	7

PROVINCIA	N. ALUMNOS
Iquique	6
Cardenal Caro	5
Huasco	5
Punilla	5
Limari	5
Arica	5
Malleco	4
Petorca	4
Palena	3
Copiapó	3

PROVINCIA	N. ALUMNOS
Itata	2
Choapa	2
General Carrera	1
Tamarugal	1
Chañaral	1
Tocopilla	1
Cauquenes	1
Ranco	1

Tabla 4.3: Número de estudiantes por provincia de domicilio

COMUNA:

El atributo que muestra la comuna del domicilio presenta 169 valores distintos, cuenta con 44 estudiantes que no presentan esta información. A continuación se presentan una tabla con las 10 comunas más populares entre los estudiantes de la FCFM (66 % de los estudiantes).

COMUNA	N. ALUMNOS
Santiago	1675
Las Condes	1125
Ñuñoa	920
Providencia	716
Maipú	581
La Florida	467
La Reina	429
Peñalolén	403
Puente Alto	329
San Miguel	297

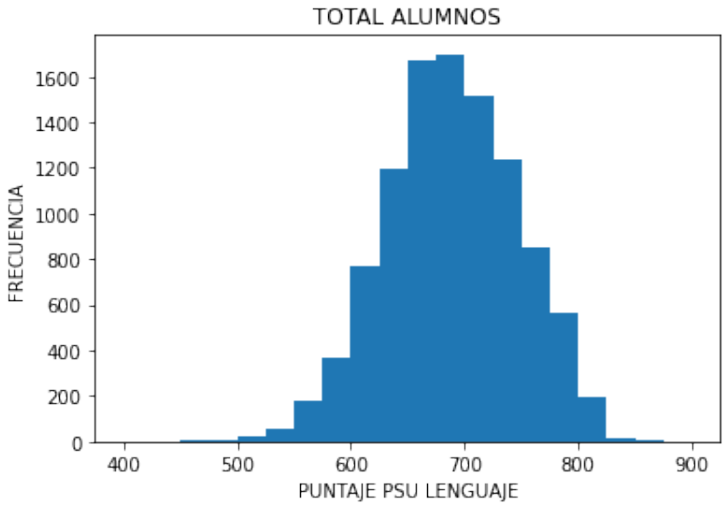
Tabla 4.4: Número de estudiantes para las 10 comunas de domicilio más recurrentes

La comuna que más se repite entre los estudiantes es la comuna de Santiago, misma comuna donde se ubica la facultad, además llama la atención que las 3 comunas siguientes corresponden a sectores donde habita población con mayores ingresos de Santiago [29].

PSU LENGUAJE:

Se realiza un histograma del puntaje PSU de lenguaje de los estudiantes. Se encuentran 56 datos perdidos, un valor mínimo de 427 y un máximo de 850. Promediando 688.6 puntos

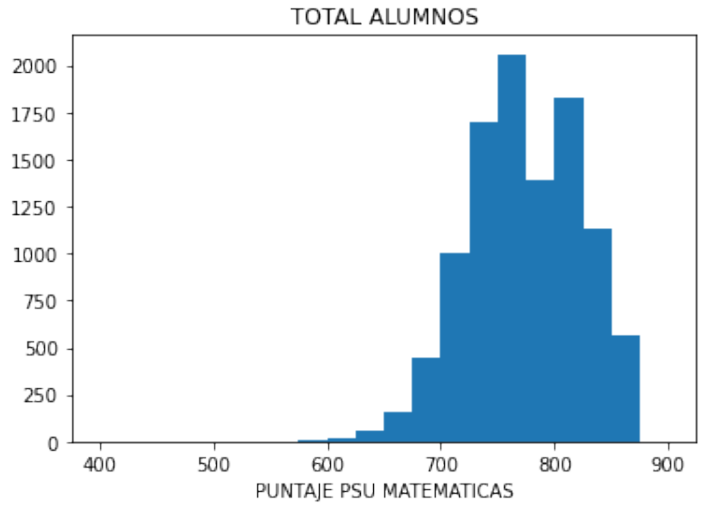
Figura 4.4: Histograma de puntajes PSU lenguaje



PSU MATEMATICAS:

Se realiza un histograma del puntaje PSU de matemáticas. Se encuentran 56 datos perdidos, un valor mínimo de 469 y un máximo de 850. Promediando 771.8 puntos. Es posible inferir la existencia de dos tipos de estudiantes, razón que explicaría la existencia de picos en el histograma.

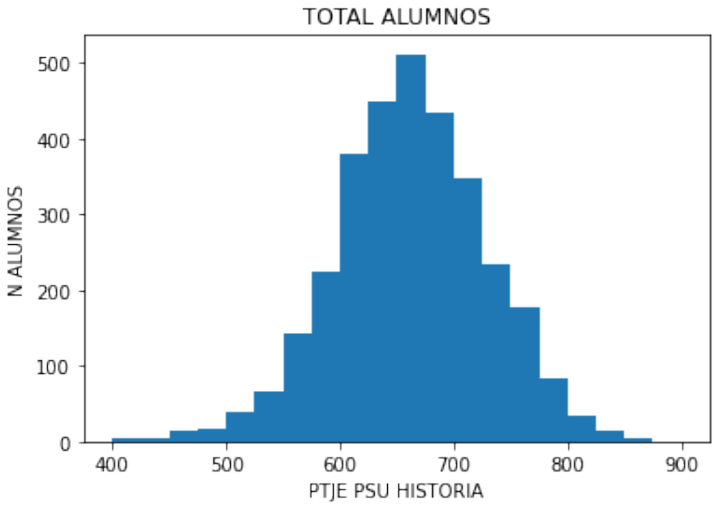
Figura 4.5: Histograma de puntajes PSU matemáticas



PSU HISTORIA:

Se realiza un histograma con los puntajes PSU de historia, se tiene un mínimo de 150 puntos y un máximo de 850 puntos con un puntaje promedio de 660.1 puntos. Se encuentran 7,220 datos faltantes, lo que significa que un 69.3 % de los estudiantes en la base de datos no presenta dicho valor. Esto se debe al carácter optativo de esta prueba.

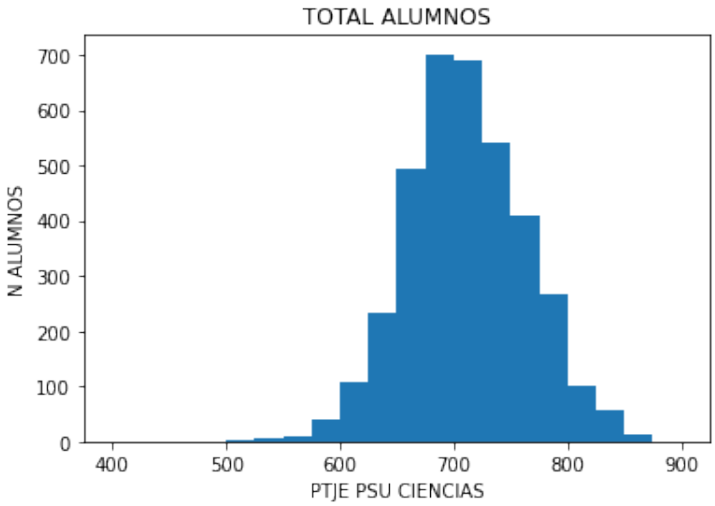
Figura 4.6: Histograma de puntajes PSU historia



PSU CIENCIAS:

El atributo captura el puntaje obtenido en uno de los 3 módulos optativos que incluye esta prueba, sin embargo, este no señala el módulo al que corresponde. Se encuentra que existen 6,736 (64.68 %) estudiantes sin este atributo. Se tiene un puntaje mínimo de 477 y un máximo de 850 puntos. Con un promedio de 710.83 puntos.

Figura 4.7: Histograma de puntajes PSU ciencias



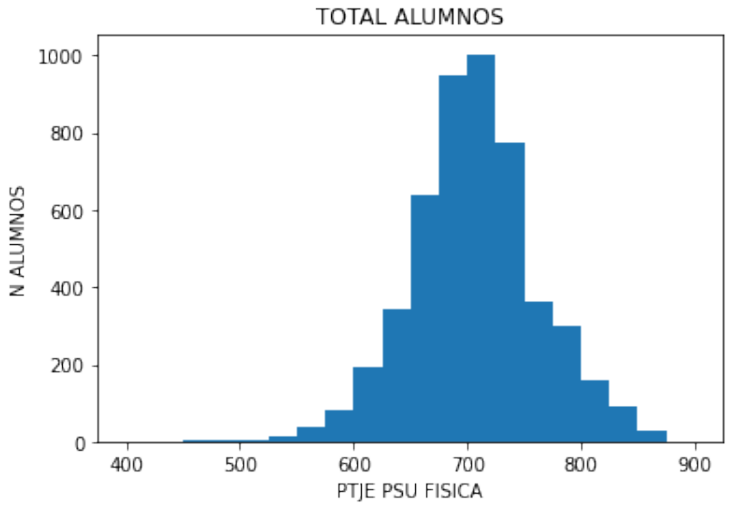
La base de datos cuenta con 3 atributos distintos que capturan de manera independiente el valor del puntaje obtenido en cada uno de estos módulos optativos (física, química y biología).

PSU CIENCIAS FISICAS:

El atributo que captura el puntaje obtenido en la prueba de ciencias, módulo física, posee 5,419 (52.04 %) datos faltantes. Toma valores que van de los 430 puntos a un valor máximo

de 850. Toman un valor promedio de 705.74 puntos.

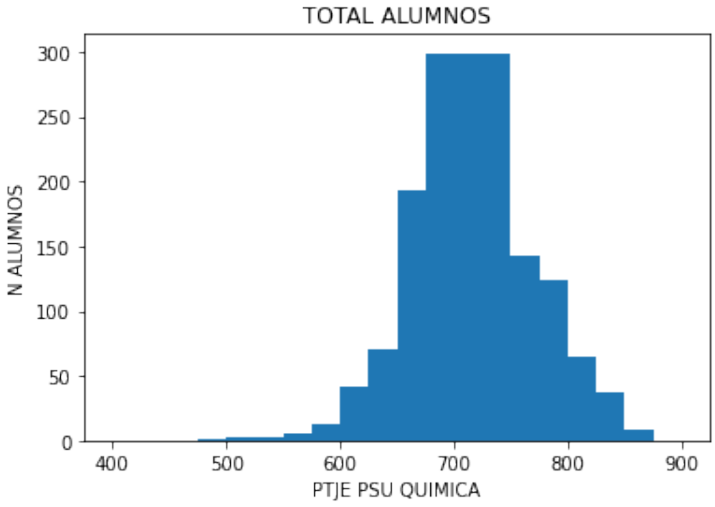
Figura 4.8: Histograma de puntajes PSU ciencias mod. física



PSU CIENCIAS QUÍMICA:

El atributo que captura el puntaje obtenido en la prueba de ciencias, módulo química, posee 8,811 (84.61 %) datos faltantes. Toma valores que van de los 496 puntos a un valor máximo de 850. Posee una distribución asimétrica positiva torno a su valor promedio de 716.14 puntos.

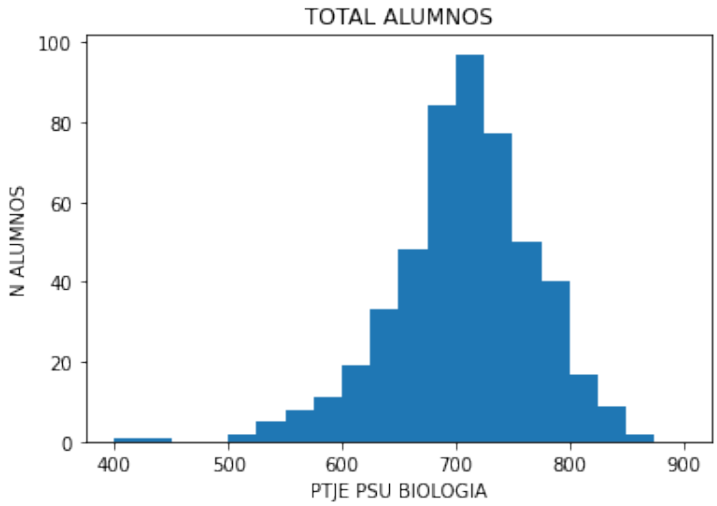
Figura 4.9: Histograma de puntajes PSU ciencias mod. química



PSU CIENCIAS BIOLOGÍA:

El atributo que captura el puntaje obtenido en la prueba de ciencias, módulo física, posee 9,909 (95.15 %) datos faltantes. Toma valores que van de los 424 puntos a un valor máximo de 850. Toman un valor promedio de 706.53 puntos

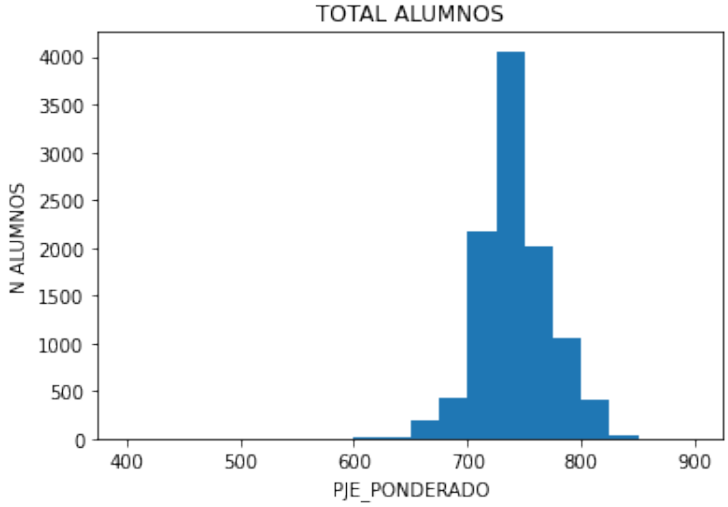
Figura 4.10: Histograma de puntajes PSU ciencias mod. biología



PUNTAJE PONDERADO:

El puntaje ponderado corresponde al puntaje con el que el estudiante postula para ingresar a la universidad, siendo seleccionados según este valor. Se presentan 4 valores atípicos [76,81,872,935]. Dejando estos fuera del análisis, se obtiene un valor mínimo de 515 y un valor máximo de 840 puntos. Con un valor promedio de 740.79 puntos.

Figura 4.11: Histograma de puntajes ponderado

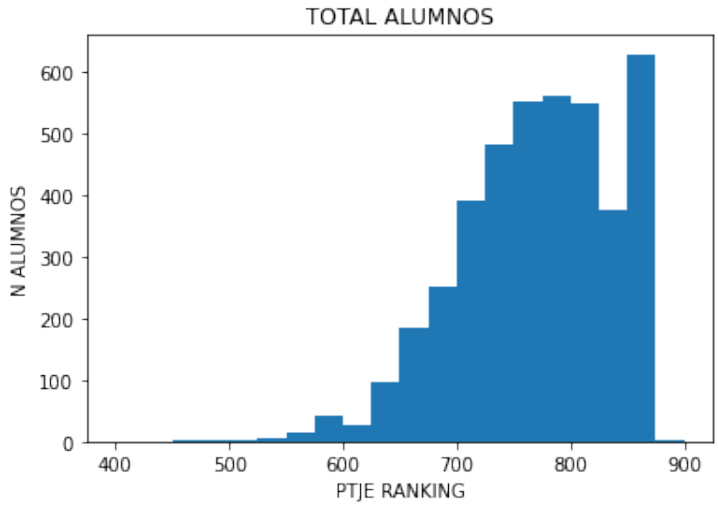


PUNTAJE RANKING:

El puntaje ranking corresponde al asignado al alumno según la posición de notas con respecto a sus compañeros en la educación secundaria. Se presentan 2 valores atípicos [881,974]. Excluyendo estos del análisis, se obtiene un valor mínimo de 269 y un valor máximo de 850 puntos. Con un valor promedio de 768.31 puntos, por último se identifican 6236 (59.88 %) valores faltantes.

La distribución se presenta con dos picos, siendo alta la cantidad de alumnos que se presentan con puntajes cercanos a los 850 puntos.

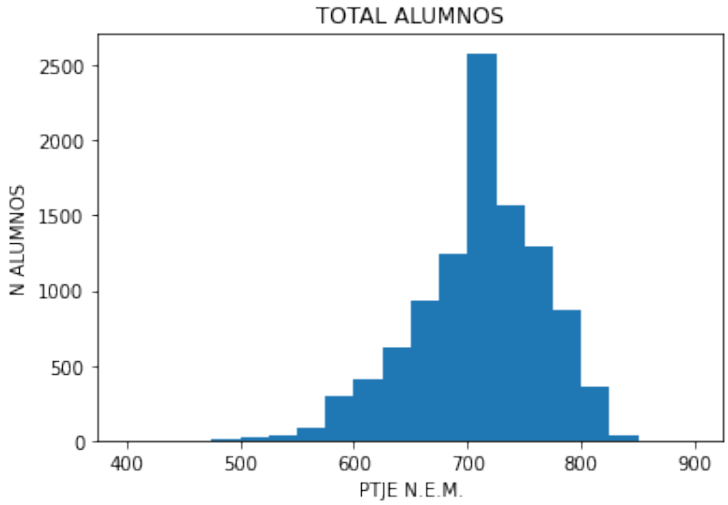
Figura 4.12: Histograma de puntajes ranking



PUNTAJE N.E.M.:

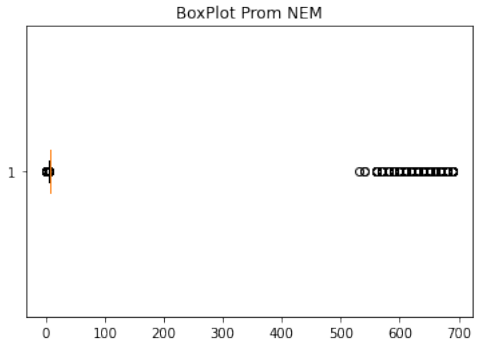
El puntaje N.E.M. corresponde al puntaje otorgado según el promedio de notas obtenido durante su enseñanza media. Se encuentran 43 datos faltantes. El atributo toma un valor mínimo de 393 puntos y un máximo de 826, se tiene un promedio de 710.81 puntos.

Figura 4.13: Histograma de puntajes N.E.M.

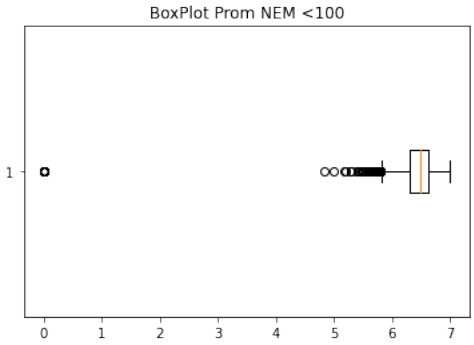


PROMEDIO NOTA ENSEÑANZA MEDIA:

Este valor corresponde al promedio obtenido por el alumno durante sus últimos cuatros años correspondientes a la enseñanza media. Se encuentran valores atípicos, valores superiores a 100, además de valores iguales a 0. Donde la escala de notas va desde 1 a 7.



(a) valores registrados

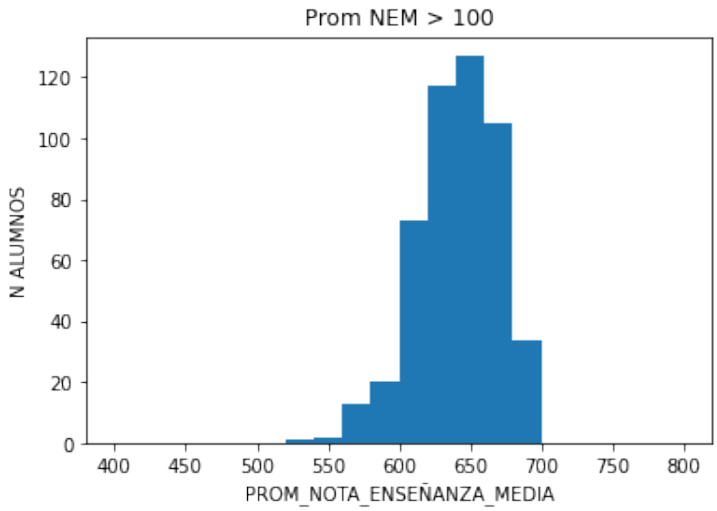


(b) valores registrados menor a 100

Figura 4.14: BoxPlot de promedio N.E.M.

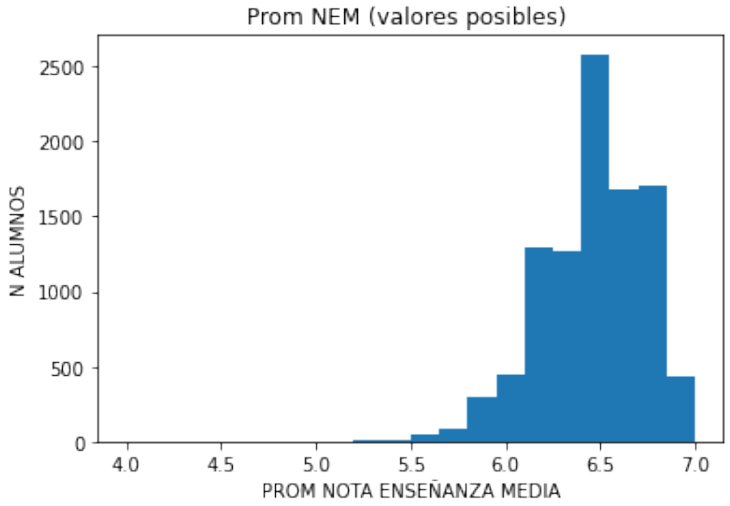
Este error proviene al ingresar el valor sin la puntuación del decimal.

Figura 4.15: Histograma de promedios de nota de enseñanza media



Excluyendo los datos atípicos, el atributo toma un valor mínimo de 4.83 y un máximo de 7.0. Con un valor promedio de 6.44.

Figura 4.16: Histograma de promedios de nota de enseñanza media



VÍA DE INGRESO:

Existen distintas modalidades de ingreso para los estudiantes, donde es posible identificar qué los dos primeros con mayor cantidad de alumnos corresponden a los medios de ingreso oficial del cohorte correspondiente.

MEDIO INGRESO	N. ALUMNOS
PSU	9856
PAA	533
BEA	384
PEG	187
SIPEE	130
DEPORTISTA	63
EXTRANJERO	22
PACE	14

PREFERENCIA

Dentro del proceso de ingreso a la universidad, los estudiantes deben ordenar sus opciones de ingreso según sus preferencias. Se observa que 8,343 (83.2 %) estudiantes optan por considerar el ingreso a ingeniería o licenciatura en la FCFM de la Universidad de Chile como su primera opción.

PREFERENCIA	N. ALUMNOS
1ra	8664
2da	1584
3ra	123
4ta	42

Tabla 4.5: Preferencia de ingreso a la FCFM

DEPENDENCIA EDUCACIONAL

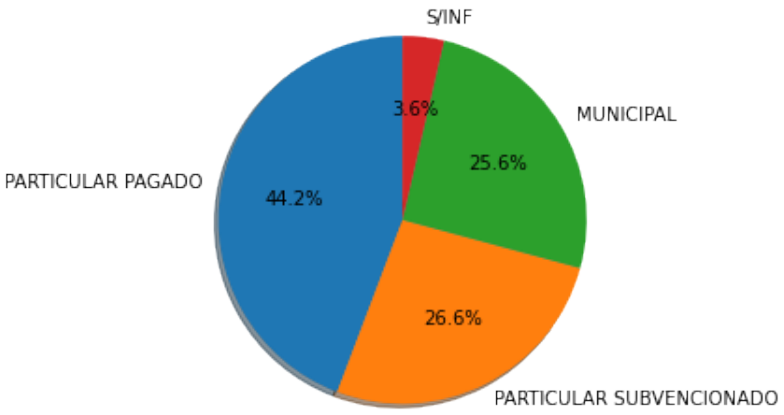
Los establecimientos de educación reconocidos por el estado son clasificados según su dependencia administrativa y financiera, tomando 3 valores dentro de la base de datos.

Municipal: Establecimiento públicos de propiedad y financiamiento principalmente estatal, administrado por la municipalidad correspondiente a la ubicación del establecimiento.

Particular Subvencionado: Establecimientos de propiedad y administración privada, sin embargo, estos reciben financiamiento estatal mediante concepto de subvención por alumno matriculado que efectivamente asiste a clases.

Particular Pagado: Establecimientos de propiedad, administración y financiamiento privado, donde este último es proporcionado por particulares y apoderados del alumno.

Figura 4.17: Distribución de dependencia educacional



Se observa qué más del 70 % de los estudiantes proviene de establecimientos privados, con mayor relevancia en establecimientos particulares pagados. Se tiene 378 estudiantes que no presentan información.

TIPO DE EDUCACION:

La educación secundaria o de enseñanza media en Chile, está dividida en 3 grupos, Científico-Humanista, Técnico Profesional y Artística, con una duración de 4 años.

TIPO DE EDUCACIÓN	N. ALUMNOS
HUMANISTA CIENTIFICO DIURNO	8117 (77.9 %)
HUMANISTA CIENTIFICO NOCTURNO	1847 (17.7 %)
TECNICO PROFESIONAL INDUSTRIAL	160 (1.5 %)
TECNICO PROFESIONAL COMERCIAL	142 (1.3 %)
TECNICO PROFESIONAL SERVICIOS Y TECNICA	90 (0.8 %)
No Informado	37 (0.3 %)
TECNICO PROFESIONAL AGRICOLA	18 (0.1 %)
TECNICO PROFESIONAL MARITIMA	2 (0.0 %)

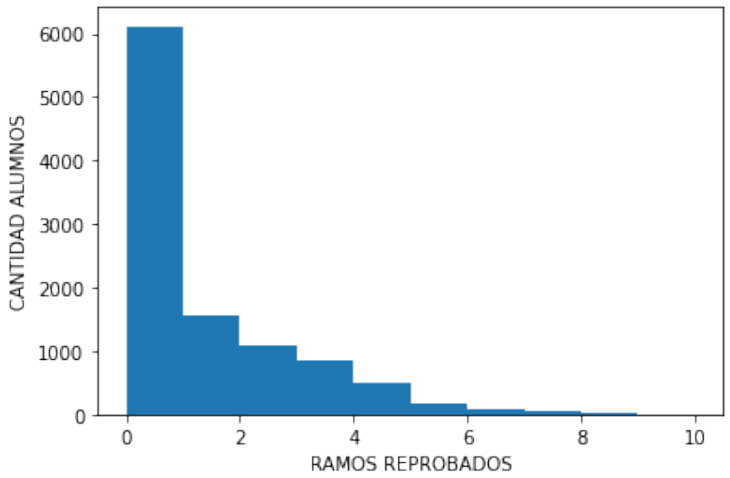
Tabla 4.6: Distribución por tipo de educación

Más del 95 % de los estudiantes dentro de la base de datos recibió educación del tipo Humanista Científico, siendo en mayor proporción de modalidad diurna. Por otra parte, se tiene un número muy inferior de estudiantes provenientes de educación Técnica Profesional.

RAMOS REPROBADOS:

Este atributo corresponde al total de ramos que reprueba el alumno durante su primer año de universidad. El atributo cuenta con un mínimo de 0 y un máximo de 9.

Figura 4.18: Distribución de número de ramos reprobados el primer año



Se tiene que 5,961 estudiantes (58.5 % de los datos) no reprueban ningún ramo durante su primer año. Mientras que 1,531 (15 %) reprueban 1 ramo y 1,067 (10.4 %) estudiantes reprueban dos ramos.

En general, los datos se encuentran en buenas condiciones y no se deben realizar grades ajustes antes de ser utilizados en los modelos de clasificación, de todas formas en la siguiente sección se realizan los ajustes necesarios a la base de datos.

4.3.1. Caracterización del Estudiantado

Los resultados expuestos anteriormente permiten realizar un análisis general sobre las características que presentan los estudiantes.

Se observa que la mayor parte de la población corresponden a estudiantes del género masculino, encontrándose a una razón de 3:1 con respecto a las de género femenino. La gran mayoría de estudiantes registran domicilio en la Región Metropolitana, siendo las comunas de mayores ingresos las que predominan sobre las demás, también se encuentra que la mayor proporción de estudiantes provienen de colegios privados (particulares subvencionados y particulares privados). Los datos reflejan que, en general, los estudiantes ingresan de manera adecuada en los tiempos a la institución, es decir, sin pausas significativas dentro del sistema educacional.

Los resultados obtenidos en las pruebas de selección reflejan el alto conocimiento académico al ingresar la universidad, además estos buenos resultados permiten que los estudiantes ingresen, en su mayoría, a la primera opción del proceso de selección universitaria, finalizando así en un proceso conforme a la solicitud del estudiante.

4.4. Modificación de Base de Datos

Esta etapa consiste en modificar los datos seleccionando, creando o transformando las variables a utilizar, con el objetivo de que estén listas para ser utilizadas en la etapa de modelación. La cantidad de registro para los cohortes anterior al 2003 carecen de sentido y por lo tanto se excluyen del análisis, ya que se pone en duda la validez de estos datos. Por lo tanto, solo se consideran estudiantes de cohorte 2003 en adelante para el presente trabajo investigativo.

Reparación de Datos

El análisis exploratorio permite identificar atributos con valores que carecen de sentido lógico, por lo tanto se reparan estos de tal forma que sean útiles para la investigación.

Edad de ingreso : Se tiene 501 estudiantes que presentan una edad de ingreso superior a los 100 años, careciendo de sentido lógico, por lo tanto para dicho atributo se opta por suprimir estos valores para ser reemplazados mediante una técnica de imputación.

Se utiliza un método de regresión lineal como método de imputación, utilizando el valor numérico del rut y el año de nacimiento del estudiante. La utilización de este método se basa en la idea de que la asignación de este número se realiza de manera incremental y ordenada en la medida que se registran en el Servicio de Registro Civil e Identificación de Chile. Por lo tanto, la edad de ingreso se obtiene mediante la diferencia del año de ingreso y el año de nacimiento pronosticado. Se obtiene un r2 de 0.918 para dicho modelo.

$$\text{Año de } Nacimiento = 1937{,}19 + 3{,}01\text{e} - 06 * RUT \tag{4.1}$$

$$\text{Edad de } Ingreso_{pronosticado} = \text{Año de } Ingreso - \lfloor \text{Año de } Nacimiento \rfloor \tag{4.2}$$

Promedio nota enseñanza media : Por otra parte, existen valores de promedio de

enseñanza media que se encuentran entre 540 y 690, siendo todos estos múltiplos de 10, por lo tanto se asume que estos fueron mal imputados omitiéndose el punto decimal. Esto es solucionado dividiendo por 100 el valor y reemplazándolo.

Datos Faltantes

En la etapa anterior se permite reconocer los datos faltantes, como también valores correspondientes a la ausencia de información. Con el fin de dimensionar este defecto sobre la base de datos se expone la siguiente tabla.

ATRIBUTO	DATOS FALTANTES
GENERO	0.02 %
TIPO EDUCACION	0.36 %
PTJE NEM	0.42 %
REGION	0.49 %
PTJE PSU LENGUAJE	0.54 %
PTJE PSU MATEMATICA	0.54 %
PROM NOTA ENSEÑANZA MEDIA	0.55 %
DEPENDENCIA EDUC	3.65 %
PTJE PSU CIENCIAS FISICA	51.95 %
PTJE RANKING	59.67 %
PTJE PSU CIENCIAS	64.76 %
PTJE PSU HISTORIA CS	69.56 %
PTJE PSU CIENCIAS QUIMICA	84.52 %
PTJE PSU CIENCIAS BIOLOGIA	95.18 %

Tabla 4.7: Porcentaje de datos faltantes por atributo

Existen atributos que presentan un bajo porcentaje de datos faltantes, para estos simplemente se opta por suprimir estos registros de la base de datos, estos corresponden a datos faltantes en los atributos de género, tipo de educación, puntaje NEM, región, PSU de lenguaje, PSU de matemáticas, promedio enseñanza media y dependencia educacional. La eliminación de estos registros significa una perdida del 4,58 % de los datos.

El alto número de datos faltantes para el atributo de puntaje ranking se explica debido a ser un factor incorporado al proceso de selección en el año 2013, y por lo tanto, no existen estudiantes que posean este atributo para los cohortes anteriores. Por esta razón se opta por no hacer uso de este atributo para la investigación.

Igualmente, se observa un alto número de datos faltantes para los atributos que recopilan los puntajes obtenidos en las pruebas de selección electivas y sus módulos, de la misma forma la ausencia de estos datos corresponde al proceso mismo y no una pérdida de información, por lo tanto no corresponde aplicar ninguna técnica de imputación. De esta forma, se decide eliminar el atributo de puntaje PSU de historia, además este no es considerado dentro del proceso de selección para ingresar a la FCFM, no así el puntaje asociado a las pruebas de ciencias.

Puntaje PSU Prueba Electiva

Los puntajes obtenidos en los módulos electivo de la prueba ciencias muestran un alto número de datos faltantes, por esta razón se opta por eliminar estos datos, sin embargo con el fin de no perder información se crea un único atributo que agrupe estos puntajes. A pesar de ello, dentro de la base de datos se tiene a 478 alumnos que muestran valores en mas de un módulo, hecho que no tiene sentido lógico debido que el proceso de selección solo permite realizar una de estos módulo electivos, por lo tanto se opta por eliminar los datos de estos estudiantes, siendo no considerados en la investigación.

Si bien, es posible crear un atributo categórico que corresponda al módulo electivo de la prueba de ciencias, para un número importante de estudiantes no es posible reconocer el módulo electivo (34.5 %), debido que solo muestran información en el atributo de PSU de ciencias, y no así en ninguno de los atributos correspondiente a los módulos electivos.

Ingreso Medio por Comuna

Como ya se ha señalado, diversos autores defienden la idea de que los fenómenos estudiados se ven influenciados por factores socio-económicos, sin embargo la base de datos no cuenta con atributos que reflejen directamente este factor. Por lo tanto, con el fin de obtener mayor información sobre esta dimensión se realiza un cruce de datos con datos provenientes del Observatorio Chileno de Salud Pública mediante su publicación del año 2011 [30]. De esta forma, es posible agregar a la base de datos el valor de 'Ingreso Medio por Comuna', si bien no se cuenta con mayor información actualizada, si permite tener en consideración del factor socio-económico bajo el supuesto que el ingreso medio por comuna informado en el documento y el ingreso actual no presentan grandes diferencias.

Agrupación de Variables Categóricas

Gracias al análisis exploratorio, se pude reconocer los múltiples niveles que poseen algunas variables categóricas, lo que puede ser problemático para la etapa de modelamiento. Para tratar estos atributos existen dos técnicas, una consiste en generar múltiples atributos binarios para cada nivel, o bien asignar a cada nivel un valor numérico, sin embargo ninguna de estas es apropiada debido al alto número de niveles que poseen los atributos, ademas de no existir una jerarquía entre estos niveles que permitan ser reemplazados con un valor numérico.

Por lo tanto se opta por agrupar, reduciendo el número de niveles, para después ser tratado por medio de la creación de nuevas variables binarias correspondiente al nivel.

COMUNA En general las comunas que se presentan en la base de datos corresponden a comunas de Santiago, por esta razón se utiliza una agrupación geográfica, dividiendo a las comunas de Santiago en Centro, Nororiente, Norte, Norponiente, Sur, Suroriente, Surponiente y Fuera de Santiago.

SECTOR COMUNA	N. ALUMNOS
Nororiente	3336 (35.6 %)
Norte	423 (4.5 %)
Norponiente	361 (3.9 %)
Centro	1517 (16.2 %)
Suroriente	1201 (12.8 %)
Sur	834 (8.9 %)
Surponiente	722 (7.7 %)
Fuera	980 (10.5 %)

Tabla 4.8: Número de estudiantes según agrupación de comunas

ADMISION ESPECIAL El estudiante puede ingresar a la universidad, ya sea de manera regular o participando en los distintos programas de admisión que se ofrecen, esta información es recopilada por múltiples variables binarias donde se señala si el estudiante participa o no de dichos programas. Por lo tanto, se opta por generar solo una variable que identifique si este hace ingreso de manera regular o especial. Los detalles se presentan en la siguiente tabla.

VAR. BINARIA	PROGRAMA	ADMISIÓN ESPECIAL
PSU	PRUEBA DE SELECCIÓN UNIVERSITARIA	NO
PAA	PRUEBA DE APTITUD ACADEMICA	NO
BEA	BECA EXCELENCIA ACADEMICA	SI
PEG	PROGRAMA EQUIDAD DE GENERO	SI
DEPORTISTA	INGRESO DEPORTISTA DESTACADO	SI
EXTRANJERO	ESTUDIOS MEDIOS EN EL EXTRANJERO	SI
PACE	PROG. ACOMP. Y ACCESO EFECTIVO E.S.	SI
SIPEE	S. ING. PRIORITARIO DE EQUIDAD EDUC.	SI

Tabla 4.9: Agrupación para conformar variable binaria de admisión especial

TIPO DE EDUCACIÓN Si bien el atributo logra rescatar los diferentes tipos de educación que recibieron los alumnos en su etapa educacional previa, se observa que los datos se concentran solo en las categorías de educación humanista científica, ya sea en su modalidad diurna o nocturna, mientras que el restante se distribuye entre las 5 categorías de educación técnica registradas, por lo tanto, se opta por agrupar estos en una sola categoría para no disipar su efecto al analizar los datos, siendo todas estas agrupadas bajo el valor de 'Educación Técnica'.

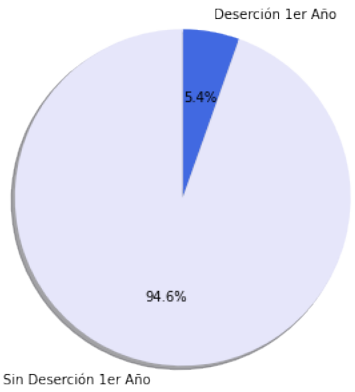
4.5. Análisis Exploratorio Sobre Casos de Estudio

Una vez completada la etapa de modificación, se realiza un análisis exploratorio utilizando cada uno de los atributos, esto con el fin de reconocer tempranamente relaciones entre los atributos y los distintos casos de estudio, para esto se procede a realizar una comparación gráfica entre estudiantes identificados como pertenecientes a la clase positiva frente a quienes pertenecen a la clase negativa según el caso de estudio.

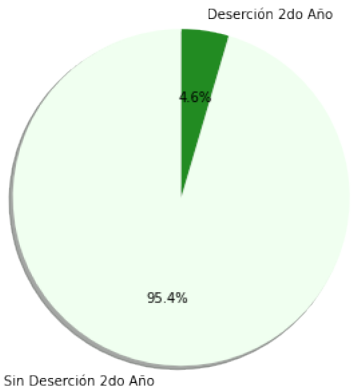
DISTRIBUCIÓN DE CASOS DE ESTUDIO

Utilizando la base de datos es posible identificar estudiantes que desertan luego del primer año, del segundo año y quienes finalizan sus estudios oportunamente, sin embargo al estudiar cada uno de estos fenómenos se observa que la distribución de clases se encuentra desequilibrada.

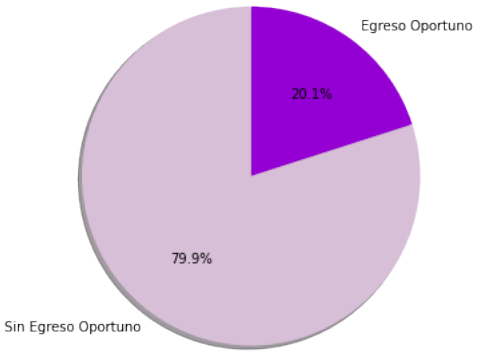
Se cuenta con 491 (5.4 %) estudiantes desertores al primer año y 8,520 (94.6 %) estudiantes que no desertan, 354 (4.6 %) estudiantes que desertan al segundo año y 7,413 (95.4 %) que continúan, y por último, 1141 (20.1 %) estudiantes que finalizan oportunamente mientras que 4,544 (79.9 %) quienes no. Se deben tratar los datos conociendo esta información para no cometer errores al entrenar los modelos de clasificación.



(a) Deserción Primer Año



(b) Deserción Segundo Año

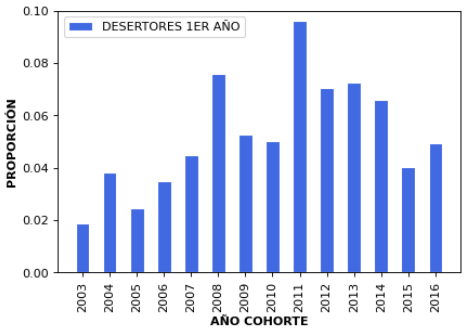


(c) Egreso Oportuno

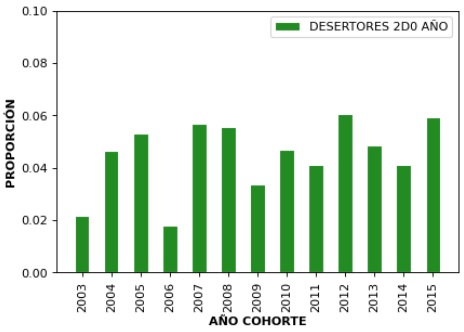
Figura 4.19: Distribución de ramos reprobados durante el primer año para cada caso de estudio

COHORTE:

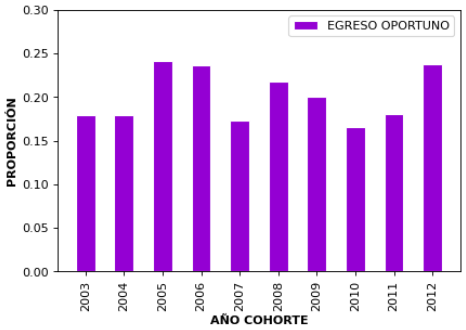
Dentro del estudio no se considera el año de cohorte como un atributo predictor para los modelos de clasificación, ya que carece de sentido realizar una predicción futura utilizando como predictor un año de ingreso posterior a los presentes en la investigación. Sin embargo, de igual manera se muestra el comportamiento de los fenómenos según el año de cohorte.



(a) Des. primer Año



(b) Des. segundo Año



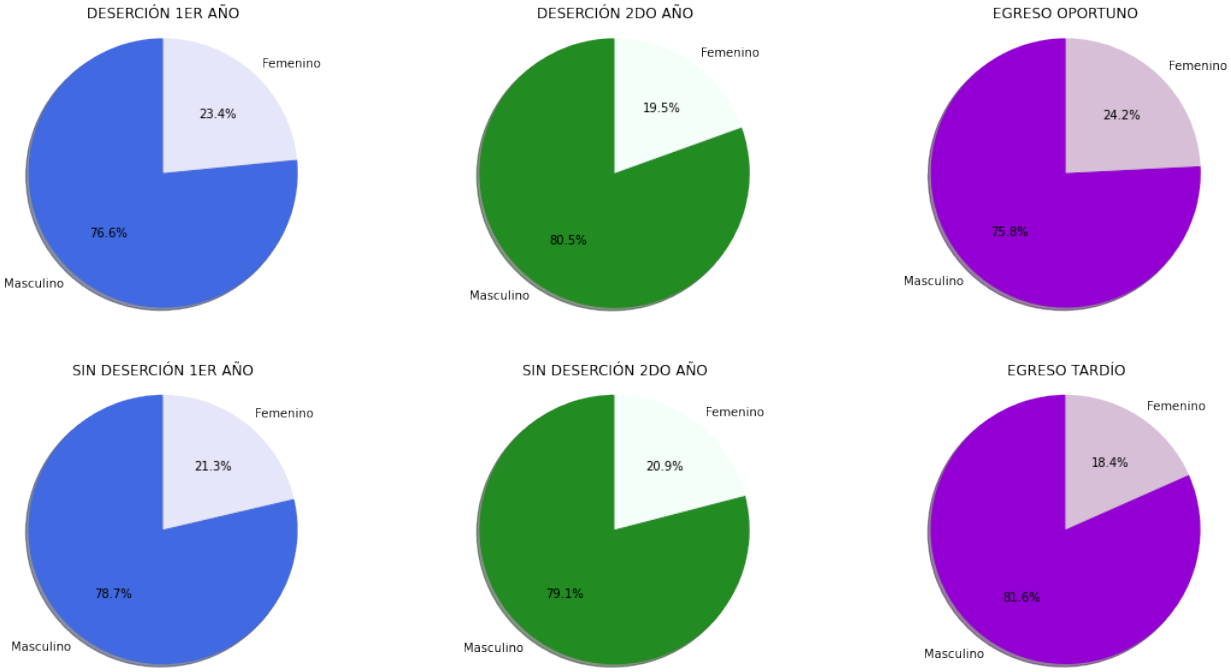
(c) Egr. Oportuno

Figura 4.20: Porcentaje de estudiantes según caso de estudio para distintos cohortes

Se encuentra que el fenómeno de deserción al primer año se presenta con mayor variabilidad entre los distintos cohortes, a diferencia de los demás casos de estudio. En particular llama la atención lo ocurrido en el año 2011, año que coincide con fuertes manifestaciones estudiantiles.

GENERO:

Se realiza una representación gráfica sobre la distribución del género en cada clase dependiendo del caso de estudio, así se observa que para los casos de deserción temprana no existen grandes diferencias entre quienes desertan y quienes no lo hacen, sin embargo existe una leve diferencia al observar a los estudiantes que egresan oportunamente frente a quienes no, donde estas diferencias de proporción alcanzan 5.8 puntos porcentuales.

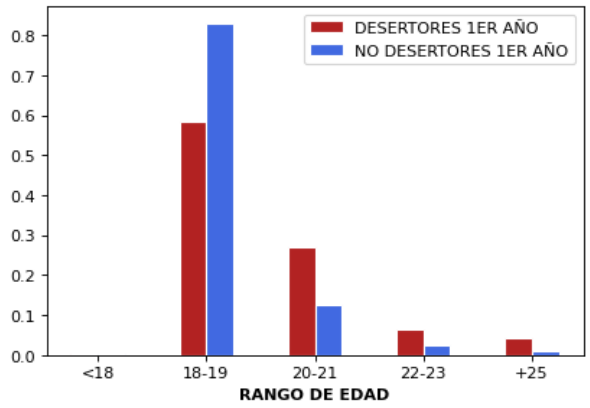


(a) Des. primer Año (b) Des. segundo Año (c) Egr. Oportuno

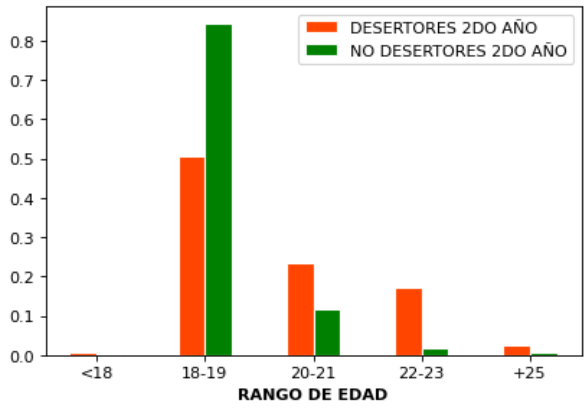
Figura 4.21: Distribución de género según caso de estudio

EDAD DE INGRESO:

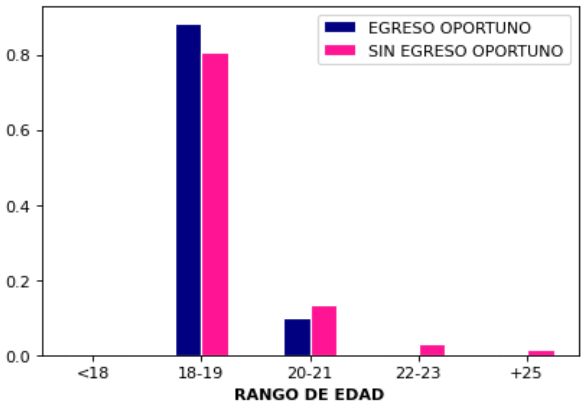
Al agrupar la edad de ingreso por rango, se observan notables diferencias entre las distribuciones. Así, se ve que quienes desertan de manera temprana presentan una mayor concentración en los rangos de mayor edad. Por otra parte, al observar el fenómeno de egreso oportuno, pese a que no hay una diferencia considerable entre las clases, se observa que quienes egresan oportunamente muestran mayor concentración en los rangos de menor edad. Por lo tanto, se puede inferir una relación entre ingresar tardíamente a la universidad y una disposición a presentar dificultades dentro del progreso de su carrera universitaria.



(a) Deserción Primer Año



(b) Deserción Segundo Año

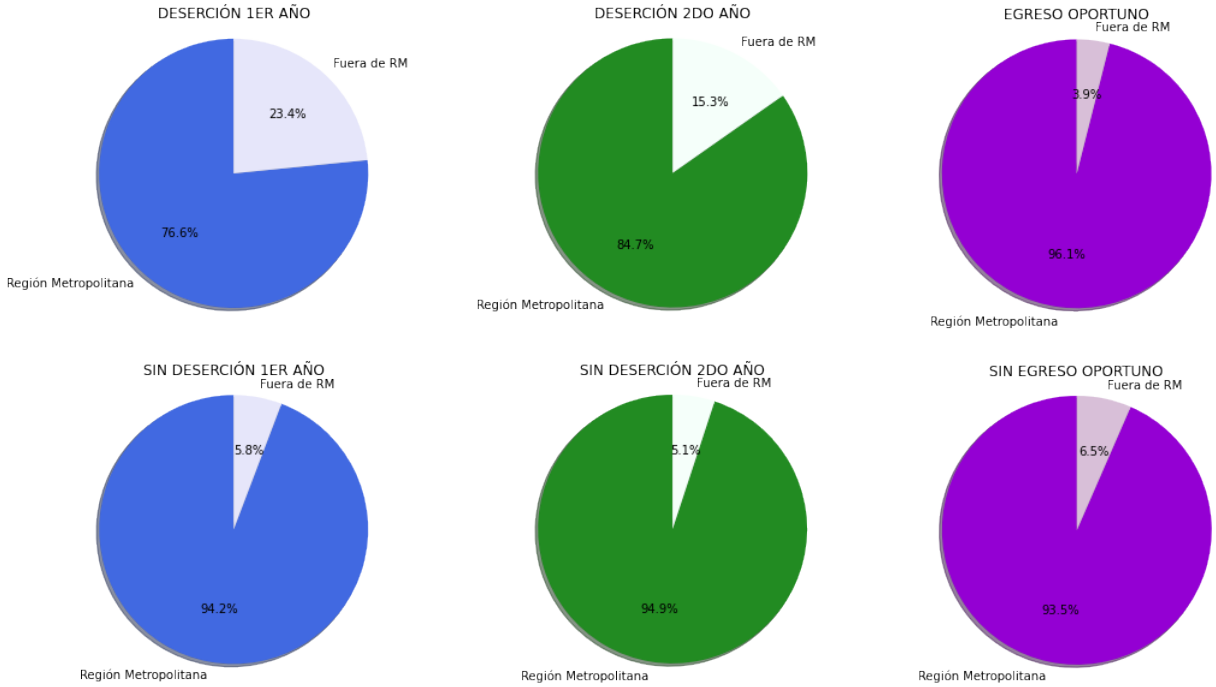


(c) Egreso Oportuno

Figura 4.22: Distribución de edad por rango según caso de estudio

REGIÓN:

Debido a la alta concentración de estudiantes con domicilio en la región metropolitana frente quienes registran otras regiones se opta por visualizar la distribución entre clases agrupando las regiones en dos grupos, Región Metropolitana o fuera de esta. Así, se observa que para los casos de deserción al primer año y deserción al segundo año, se presentan diferencias de 17.6 y 10.2 puntos porcentuales respectivamente. Mientras que, quienes egresan oportunamente presentan una mayor proporción de estudiantes de la Región Metropolitana frente a quienes no egresan oportunamente, con una diferencia menor de 2.6 puntos porcentuales. Por lo tanto, se puede inferir que los estudiantes que registran domicilio fuera de la Región Metropolitana presentan predisposición a enfrentarse a dificultades tempranamente en el progreso de su carrera universitaria.

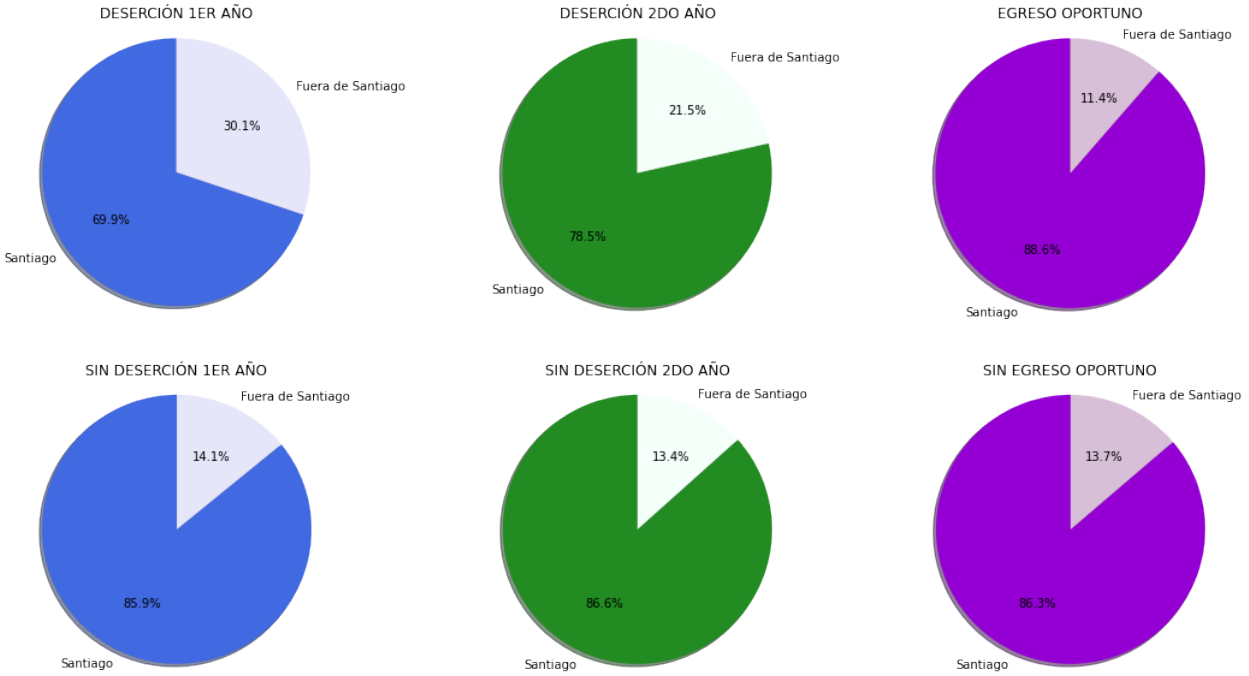


(a) Deserción Primer Año (b) Deserción Segundo Año (c) Egreso Oportuno

Figura 4.23: Distribución región según caso de estudio

PROVINCIA:

El comportamiento es similar al presentado anteriormente para el atributo de región, debido a que ambos corresponden a un atributo geográfico. De la misma forma se observa que los estudiantes fuera de la provincia de Santiago se encuentran en mayor predisposición a presentar dificultades.

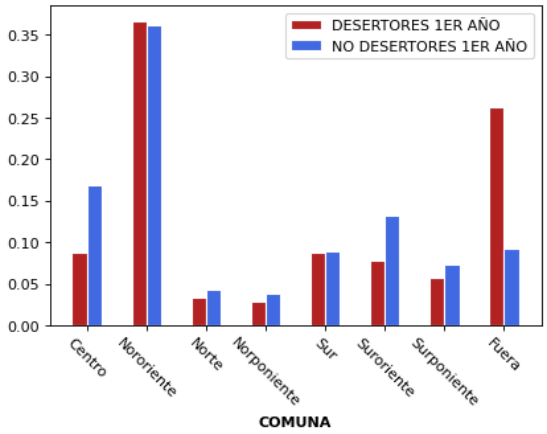


(a) Deserción Primer Año (b) Deserción Segundo Año (c) Egreso Oportuno

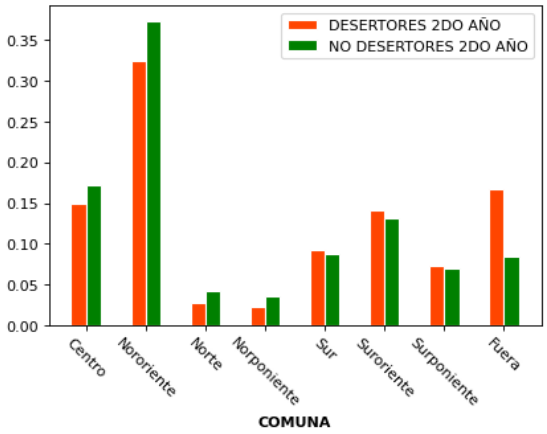
Figura 4.24: Distribución de provincia según caso de estudio

COMUNA:

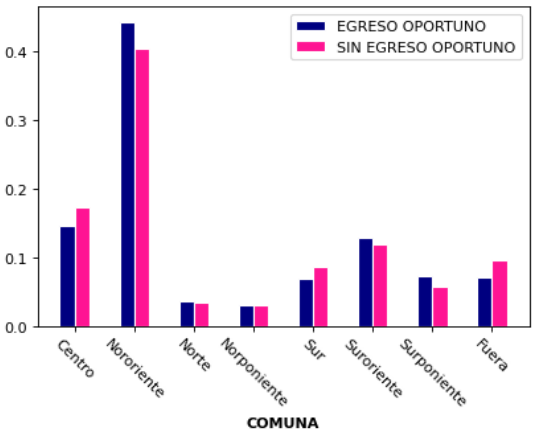
Para el caso de la deserción al primer año se tiene que las comunas pertenecientes a Santiago centro y sur-poniente presentan mejores resultados, es decir, que la proporción de estudiantes que desertan al primer año es inferior a quienes no lo hacen dentro de la comuna, mientras que el resto de sectores se mantiene relativamente constante, salvo quienes pertenecen a comunas fuera de Santiago, quienes muestran un alto incremento porcentual en el número de estudiantes que desertan. De igual forma, para la deserción al segundo año, se presenta un importante incremento porcentual sobre los estudiantes que no pertenecen a Santiago. Por último, al observar el comportamiento del atributo en estudiantes egresados, no muestran importantes diferencias entre sectores para las dos clases.



(a) Deserción Primer Año



(b) Deserción Segundo Año

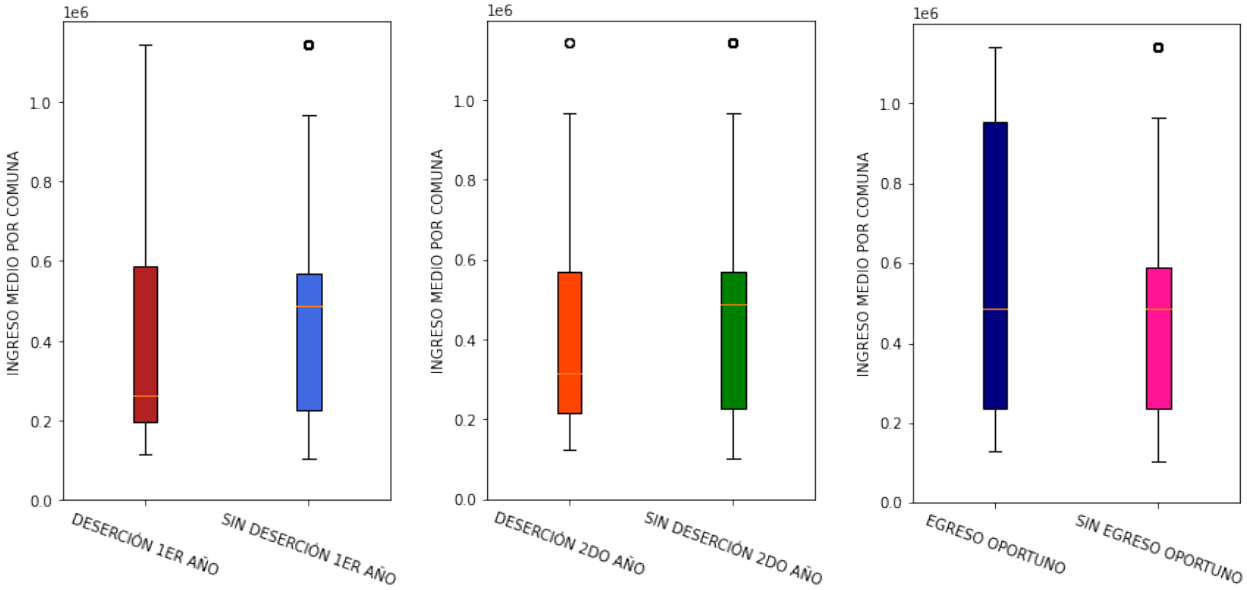


(c) Egreso Oportuno

Figura 4.25: Distribución de comuna según caso de estudio

INGRESO MEDIO POR COMUNA:

Se observa que, para ambos casos de deserción, los cuartiles presentan valores y comportamientos similares, donde el segundo cuartil, es decir la mediana de la distribución, es notablemente superior dentro de los estudiantes que no desertan. Mientras que al analizar el caso de egreso oportuno, las medianas entre los estudiantes que egresan oportunamente y quienes no, son similares. Sin embargo, no sucede lo mismo para el tercer cuartil (75 % de los estudiantes) donde en quienes egresan oportunamente este valor es cercano a los $900.000 mientras que en quienes no egresan de manera oportuna es cercano a los $600.000.

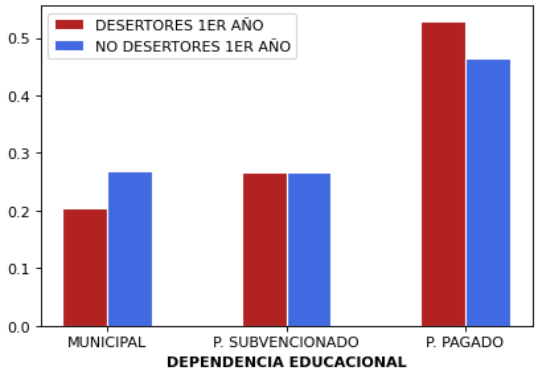


(a) Deserción Primer Año (b) Deserción Segundo Año (c) Egreso Oportuno

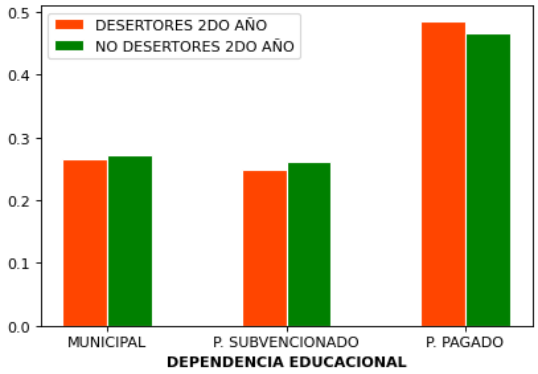
Figura 4.26: Diagrama de caja para ingreso medio por comuna para cada caso de estudio

DEPENDENCIA EDUCACIONAL:

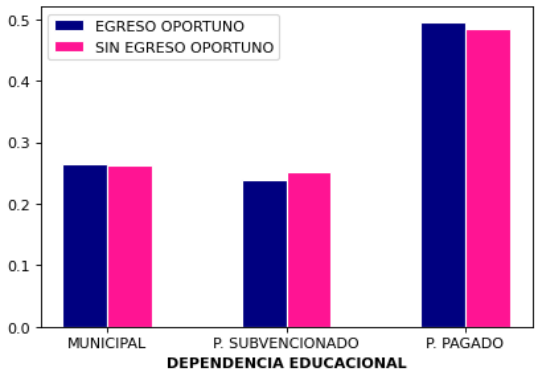
Al analizar la deserción del primer año, la proporción de estudiantes que desertan es menor en establecimientos municipales, caso contrario, ocurre en los establecimientos particular pagado, donde la proporción es mayor. Mientras que, para los casos de deserción al segundo año y egreso oportuno se tiene que la distribución de la dependencia educacional es similar tanto para los estudiantes de clase positiva como los de clase negativa.



(a) Deserción Primer Año



(b) Deserción Segundo Año

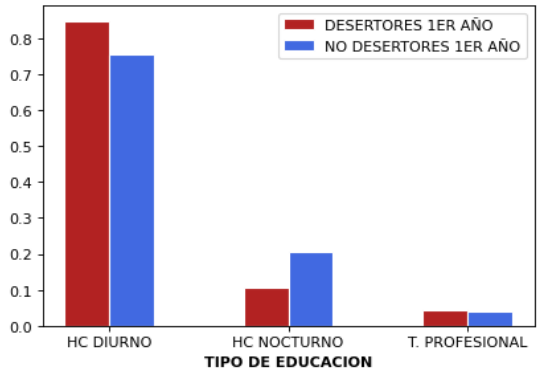


(c) Egreso Oportuno

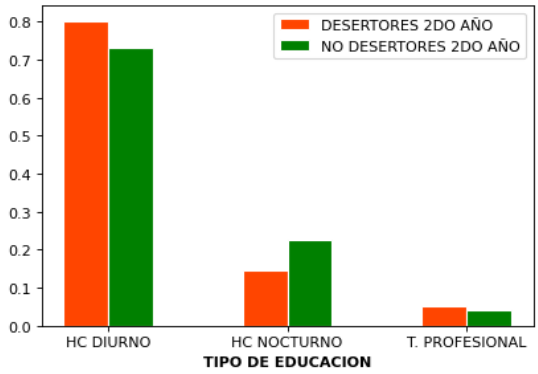
Figura 4.27: Distribución de dependencia educacional según caso de estudio

TIPO DE EDUCACIÓN:

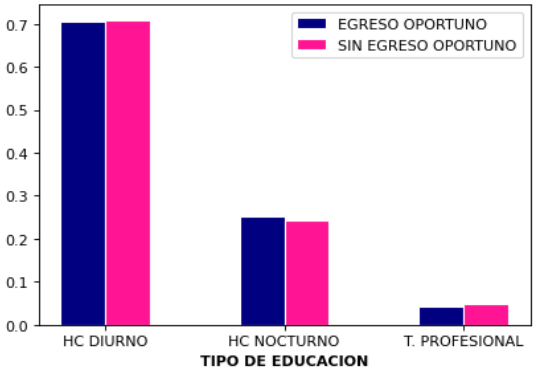
Al observar los gráficos correspondientes al estudio de la deserción, se tiene que en ambos casos la proporción de estudiantes provenientes de una educación técnica se mantiene similar, sin embargo esto no ocurre con los provenientes de una educación humanista, para ambos casos de deserción se tiene que la proporción de estudiantes desertores provenientes de la modalidad nocturna es menor con respecto a quienes no desertan, caso contrario ocurre para la modalidad diurna. Para el análisis de egreso todas las proporciones se mantienen con niveles similares.



(a) Deserción Primer Año



(b) Deserción Segundo Año

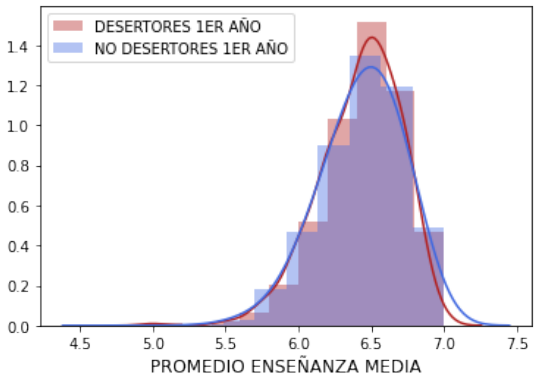


(c) Egreso Oportuno

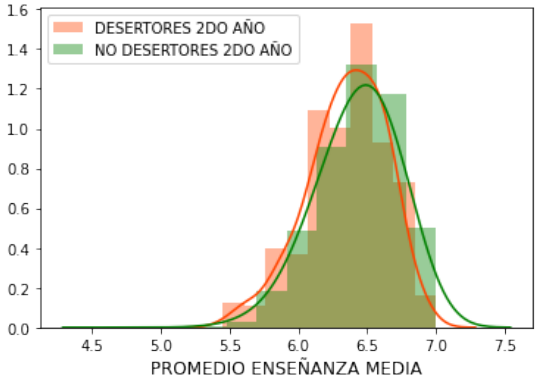
Figura 4.28: Distribución de tipo de educación según caso de estudio

PROMEDIO ENSEÑANZA MEDIA:

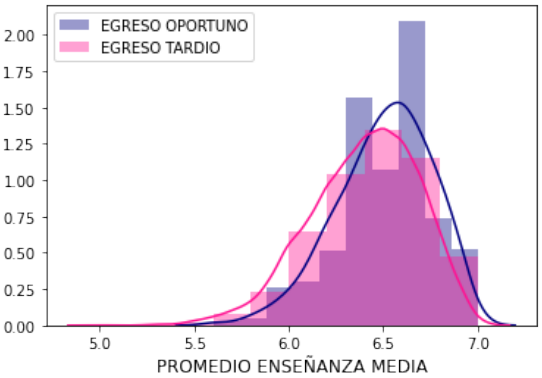
Los gráficos correspondientes a quienes desertan frente a quienes no, no difieren de manera relevante para ninguno de los dos casos de deserción. En cambio, para los estudiantes que egresan oportunamente la distribución se muestran hacia valores más altos frente a quienes no egresan oportunamente.



(a) Deserción Primer Año



(b) Deserción Segundo Año

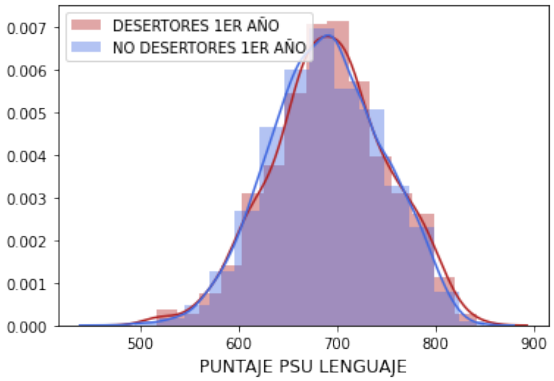


(c) Egreso Oportuno

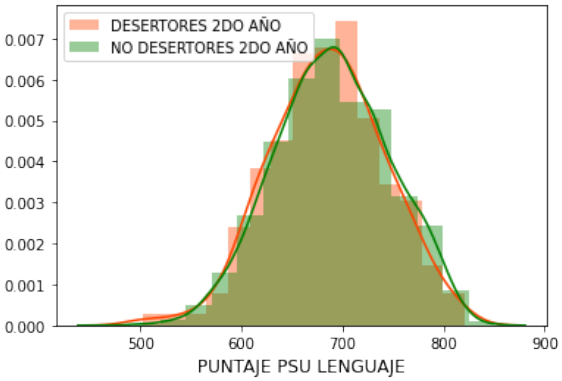
Figura 4.29: Distribución de promedio de nota enseñanza media según caso de estudio

PUNTAJE PSU DE LENGUAJE:

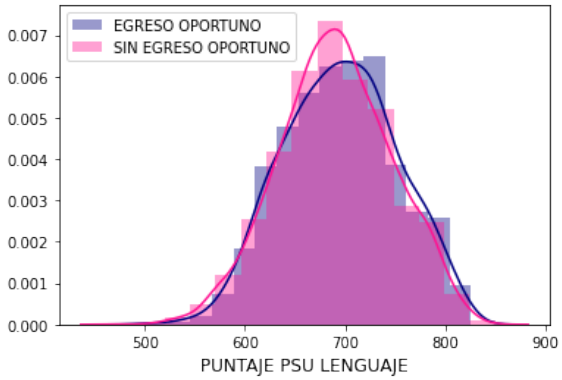
No existen diferencias apreciables al comparar el resultado de los estudiantes para los casos de deserción, mientras que quienes egresan oportunamente presentan una leve tendencia hacia mejores puntajes frente a quienes no.



(a) Deserción Primer Año



(b) Deserción Segundo Año

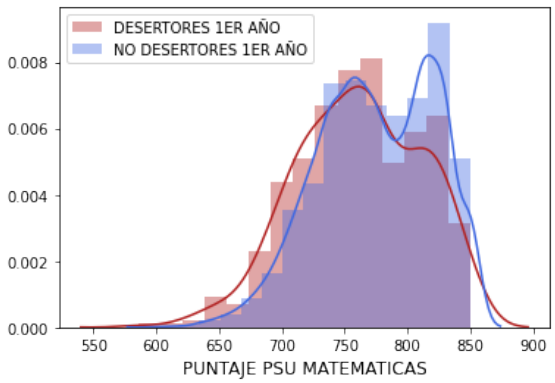


(c) Egreso Oportuno

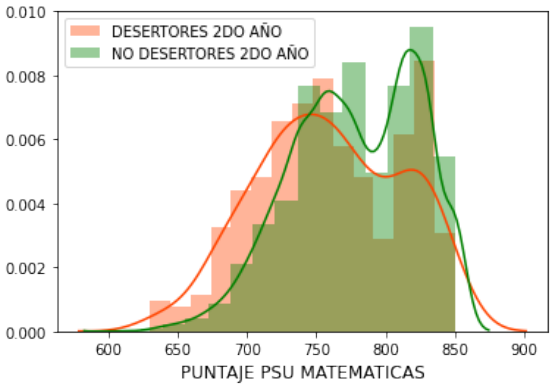
Figura 4.30: Distribución de puntaje PSU lenguaje según caso de estudio

PUNTAJE PSU DE MATEMÁTICAS:

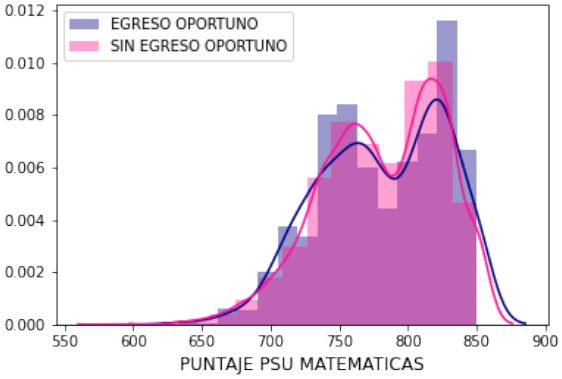
Como ya se había mencionado, la distribución de los puntajes PSU de matemáticas muestran dos picos. Pudiendo inferir la existencia de dos tipos de estudiantes. En ambos casos de deserción se replica dicho comportamiento, tanto para estudiantes desertores como no desertores, sin embargo quienes desertan muestran una acentuación mucho menor en el segundo pico, correspondientes a mayores puntajes. Para el caso de egreso oportuno no se encuentran diferencias apreciables.



(a) Deserción Primer Año



(b) Deserción Segundo Año

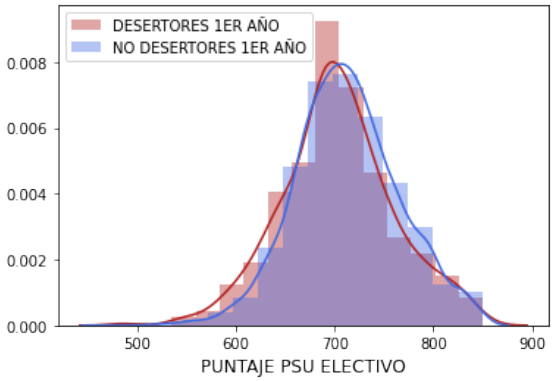


(c) Egreso Oportuno

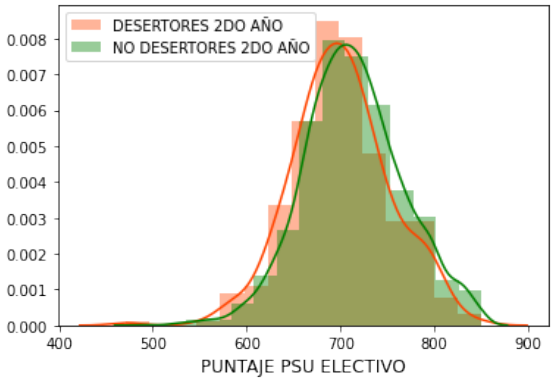
Figura 4.31: Distribución de puntaje PSU matemáticas según caso de estudio

PUNTAJE PSU ELECTIVO:

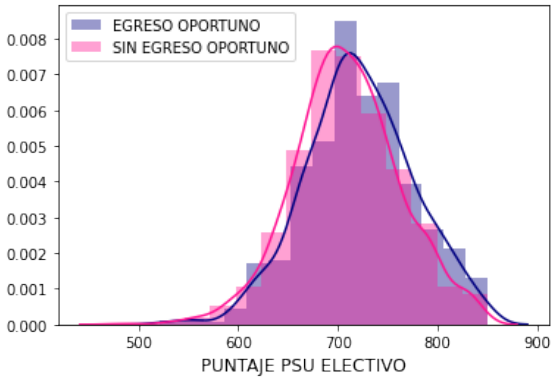
Al observar los distintos gráficos, no se presentan grandes diferencias entre la distribución de los puntajes de PSU electivo, para ninguno de los casos de estudios.



(a) Deserción Primer Año



(b) Deserción Segundo Año

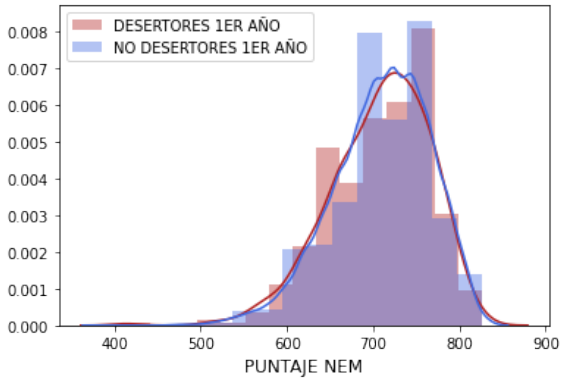


(c) Egreso Oportuno

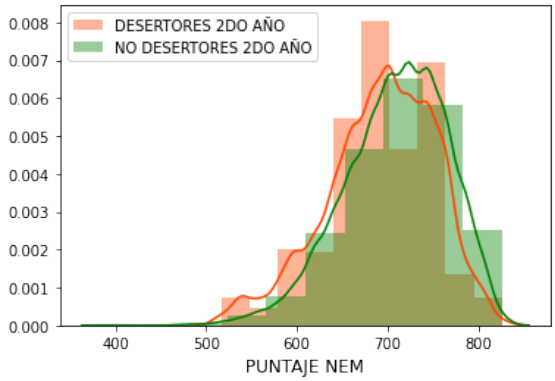
Figura 4.32: Distribución de puntaje PSU ciencias módulo electivo según caso de estudio

PUNTAJE NEM:

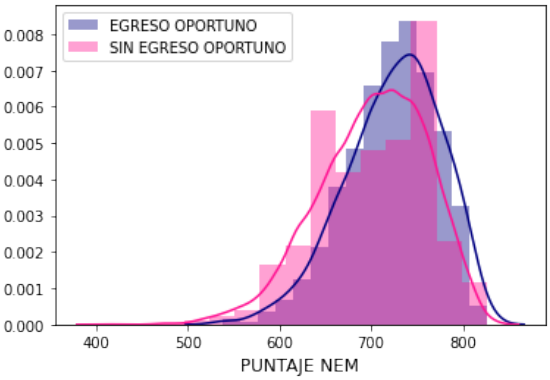
Como es de esperar, debido que este valor es calculado utilizando el promedio del estudiante en la enseñanza media, este atributo tiene un comportamiento similar al promedio de enseñanza media, antes analizado. Para los casos de deserción no existen importantes diferencias entre los estudiantes que desertan y quienes no, mientras que para los estudiantes que egresan oportunamente la distribución se concentra hacia valores más altos en comparación a quienes no egresan oportunamente.



(a) Deserción Primer Año



(b) Deserción Segundo Año

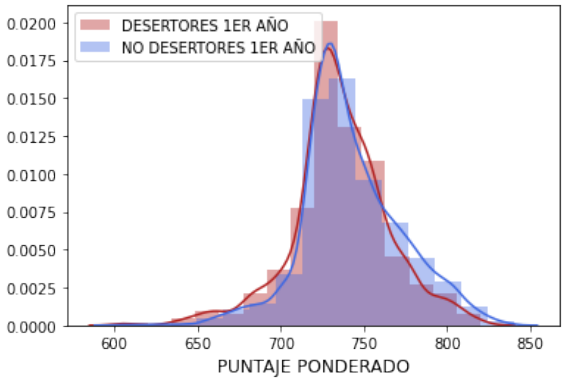


(c) Egreso Oportuno

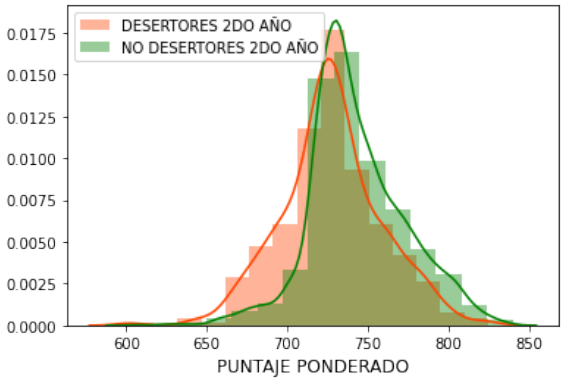
Figura 4.33: Distribución de puntaje NEM según caso de estudio

PUNTAJE PONDERADO:

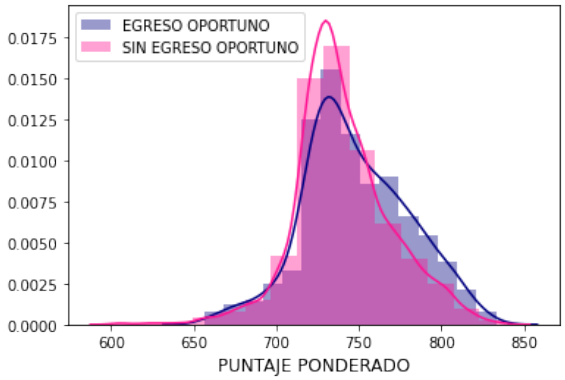
La distribución de puntajes ponderado para los estudiantes que desertan en el primer año y quienes no lo hacen no presenta diferencia aparente, mientras que los estudiantes que desertan el segundo año muestran una distribución que tiende levemente hacia menores valores, frente a quienes no desertan. Por otra parte, para quienes egresan oportunamente ocurre que la distribución de los puntajes se encuentran con leve tendencia hacia mayores valores.



(a) Deserción Primer Año



(b) Deserción Segundo Año



(c) Egreso Oportuno

Figura 4.34: Distribución de puntaje ponderado según caso de estudio

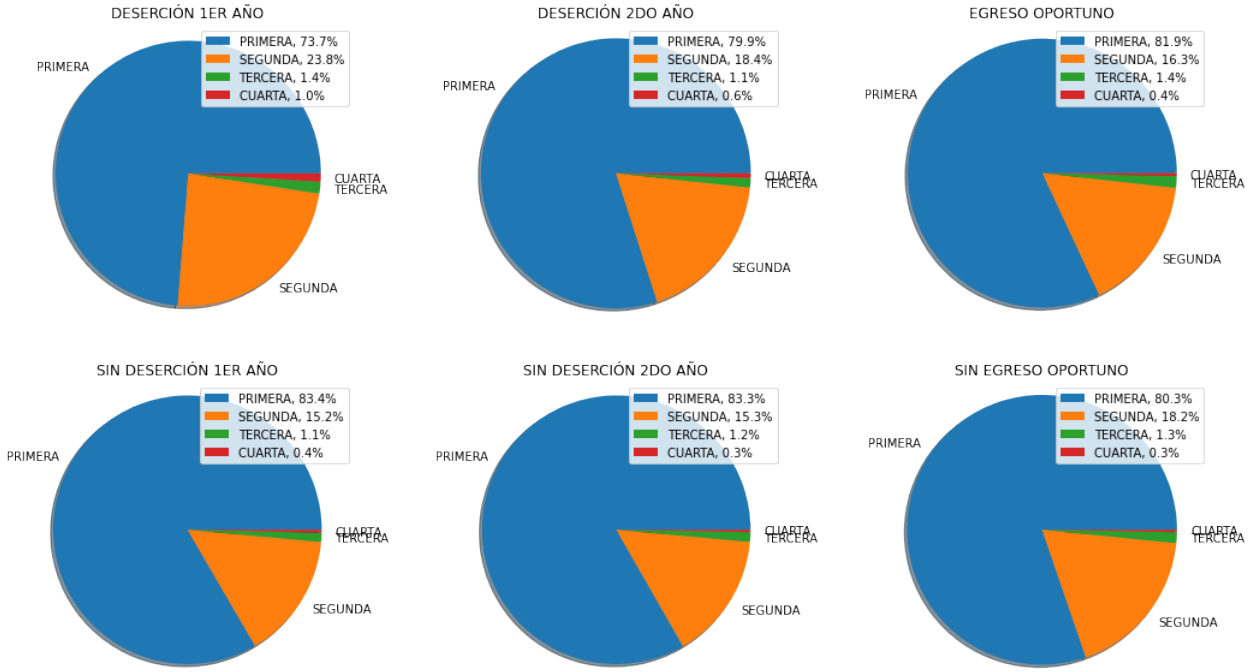
PREFERENCIA:

Se realizan comparaciones de proporción entre los grupos de clase positiva frente a quienes son de clase negativa para cada caso de estudio.

Para quienes desertan al primer año, frente a quienes no, se observa una diferencia de 9.7 puntos porcentuales dentro de los que seleccionan la carrera como primera opción y una diferencia de 8.6 puntos porcentuales en quienes la seleccionan como segunda opción. Estos mismos resultados ocurren de manera similar al analizar la deserción al segundo año, sin embargo estas diferencias son menores, siendo de 3.4 y 3.1 puntos porcentuales respectivamente.

Por lo tanto, quienes no seleccionan la carrera como primera opción se muestran más propensos a una deserción temprana.

Mientras que, al comparar las proporciones entre quienes egresan de manera oportuna y quienes no, no se aprecian diferencias relevantes.



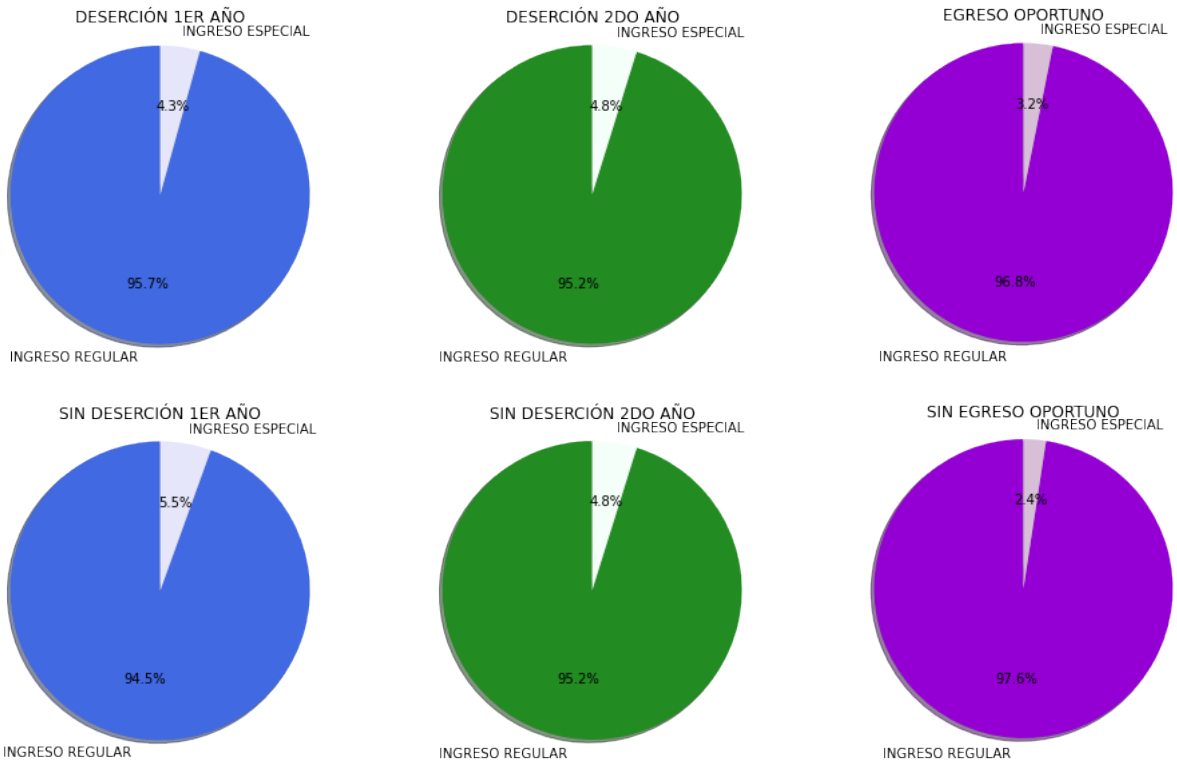
(a) Deserción Primer Año (b) Deserción Segundo Año (c) Egreso Oportuno

Figura 4.35: Gráfico de torta de preferencia según caso de estudio

ADMISIÓN ESPECIAL:

Al igual que en el caso anterior, se realiza una comparación entre las proporciones para cada caso de estudio. Las diferencias encontradas no son relevantes, estas no superan los 1.2 puntos porcentuales, incluso para el caso de estudio de deserción al segundo año ambos grupos no presentan resultados distintos.

Sin embargo, la tendencia muestra que quienes ingresan de manera especial son menos propensos a presentar las dificultades analizadas por los casos de estudio. Aún cuando las diferencias son menores.

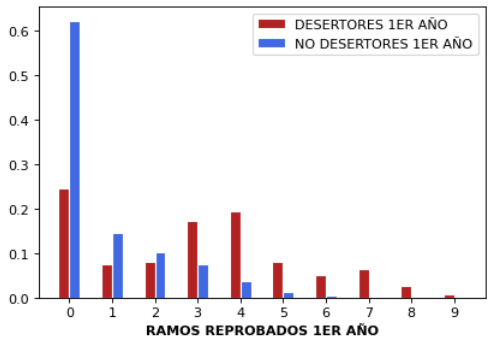


(a) Deserción Primer Año (b) Deserción Segundo Año (c) Egreso Oportuno

Figura 4.36: Distribución de estudiantes según vía de acceso para cada caso de estudio

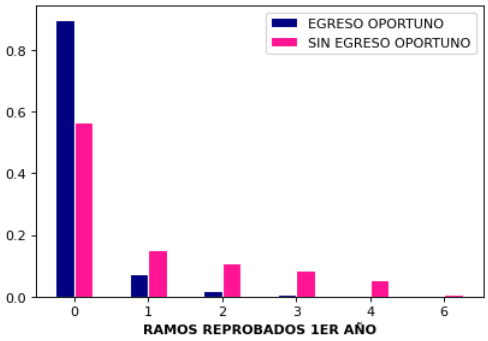
RAMOS REPROBADOS:

A continuación se presenta la distribución del número de ramos reprobados durante el primer año académico. Se observa que para quienes no desertan esta distribución se concentra torno al valor cero, alcanzando una proporción sobre del 60 % para ambos casos de deserción, mientras que para quienes desertan se observa una curva con dos picos, mostrando un incremento en las proporciones de estudiantes con mayor numero de ramos reprobados. Como es de esperar, ocurre que quienes egresan oportunamente muestran una frecuencia menor de ramos reprobados, esto es evidente debido que la reprobación de ramos impacta directamente al avance curricular del estudiante.



(a) Deserción Primer Año

(b) Deserción Segundo Año



(c) Egreso Oportuno

Figura 4.37: Distribución de ramos reprobados durante el primer año para cada caso de estudio

4.6. Algoritmos de Clasificación

El presente trabajo propone la aplicación de distintos métodos de clasificación para modelar los casos de estudio, a fin de conseguir esto, se utiliza principalmente la biblioteca Scikit-learn para Python, esta es una biblioteca gratuita de aprendizaje automático que cuenta con varios algoritmos ya programados, ademas de contar con herramientas que son útiles para la investigación. A continuación se presentan las funciones utilizadas para emplear los modelos de clasificación.

Modelo	Función
Regresión Logística	LogisticRegresion()
Árbol de Decisión	DecisionTreeClassifier()
Naive Bayes	GaussianNB()
K-Vecinos Cercanos	KNeighborsClassifier()
Redes Neuronales	MLPClassifier()
Super Vector Machine	LinearSVC()
Bagging	BaggingClassifier()
Boosted	AdaBoostClassifier()
Stacked	StackingClassifier()

Tabla 4.10: Funciones en Scikitlearn según modelo

Los últimos 3 corresponden a modelos ensamblados. Estos utilizan de manera estratégica los resultados proveniente de otros modelos con el objetivo de contar con mejores predicciones.

Ajuste de Clasificadores

Los algoritmos de aprendizaje automático regularmente dependen de parámetros que determinan el modo en el que se encuentra la función óptima de clasificación, estos paramentos son conocidos como hiperparámetros del modelo.

Los hiperparámetros son valores de configuración tal que no se aprenden directamente de los datos, y por esta razón se definen antes del proceso de entrenamiento. La configuración óptima de estos no se conoce a priori y se pueden utilizar valores genéricos, valores que han funcionado bien para problemas similares o realizar una búsqueda mediante prueba y error.

En consecuencia, antes de implementar los modelos de clasificación, primero se realiza un ajuste sobre sus hiperparámetros. Para esto, se recurre a la construcción de una malla de vectores utilizando un vector diferente para cada hiperparámetro que se busca ajustar, por medio de la combinación de sus valores se compone un espacio que cuenta con todas las combinaciones posibles de valores. Esta técnica es conocida como búsqueda en malla y la librería Scikit-learn cuenta con dicho algoritmo a través de la función 'GridSearchCV'.

Para hacer uso de esta técnica es necesario definir una métrica de rendimiento, la que es utilizada como criterio, los hiperparámetros seleccionados son los que maximizan esta métrica.

La selección de métrica de rendimiento se hace considerando que se cuenta con bases de datos desbalanceadas, por lo tanto se opta por utilizar la métrica de f1-score como principal métrica para evaluar cada uno de los métodos a implementar, la métrica accuracy no es adecuada cuando hay desequilibrio de clases [31]. Se suma la aplicación del procedimiento de validación cruzada con 5 pliegues para obtener un valor de rendimiento con menor sesgo.

A continuación se presentan los valores seleccionados para el ajuste de hiperparámetros, donde se aplica la técnica de búsqueda en malla utilizando todos los atributos disponibles, ya que a priori no se conoce cuales de estos son los apropiados para cada caso de estudio.

MODELO DE CLASIFICACIÓN	AJUSTE DE HIPERPARÁMETROS	
	CE DESERCION 1ER AÑO	CE DESERCION 2DO AÑO
Logistic Regresion	C: 0.1 penalty: l2 solver: saga	C: 0.1 penalty: l2 solver: sag
DecisionTree	criterion:criterion: gini max_depth: 12 min_samples_leaf: 1	criterion:criterion: gini max_depth: 9 min_samples_leaf: 1
KNNeighbors	metric: manhattan n_neighbors: 3 weights: distance	metric: euclidean n_neighbors: 3 weights: distance
NeuralNet	activation: relu hidden_layer_sizes: (50,)	activation: tanh hidden_layer_sizes: (50,)
SVectorMachine	C: 100 penalty: l2	C: 100 penalty: l2

Tabla 4.11: Hiperparámetros ajustados para C.E. deserción

Se observa que los hiperparámetros seleccionados para cada modelo al estudiar la deserción son similares, esto se puede explicar debido a la similitud que existe entre estos fenómenos, además de estar usando casi los mismos atributos para realizar el ajuste (para el caso de deserción al segundo año, se cuenta con el rendimiento académico).

MODELO DE CLASIFICACIÓN	AJUSTE DE HIPERPARÁMETROS
	EGRESO OPORTUNO
Logistic Regresion	C: 0.001 penalty: l2 solver: liblinear
DecisionTree	criterion:criterion: gini max_depth: 20 min_samples_leaf: 1
KNNeighbors	metric: manhattan n_neighbors: 3 weights: distance
NeuralNet	activation: tanh hidden_layer_sizes: (50,)
SVectorMachine	C: 100 penalty: l2

Tabla 4.12: Hiperparámetros ajustados para C.E. egreso oportuno

Mientras que para estudiar el fenómeno de egreso oportuno, los hiperparámetros presentan mayores diferencias con respecto a los otros casos de estudio, a pesar de esto, tampoco difieren de manera significativa. Entendiéndose la similitud de estos fenómenos.

No se realiza este ajuste para el modelo de Naive Bayes debido a que no se cuenta con probabilidades previas para ninguno de los casos de estudios, siendo este el único hiperparámetro que se permite ajustar dentro de la función que se utiliza.

Método para Evaluar Clasificadores

Para evaluar el desempeño de cada modelo se procede a comparar los rendimientos en igualdad de condiciones, por lo tanto se aplica el mismo procedimiento para cada uno de estos casos. Además se realizan técnicas que permiten obtener mejores resultados, este procedimiento se detalla a continuación.

Se cuenta con bases de datos desbalanceadas, presentando un caso severo en los casos estudiados de deserción, por lo tanto se utiliza la métrica f1-score como la métrica de evaluación principal, sin embargo, también se hace la evaluación de 3 métricas secundarias que permiten abarcar otras dimensiones de evaluación, permitiendo complementar y entender de mejor manera el rendimiento real del modelo, estas son accuracy, precision y recall.

No obstante, el desbalanceo de datos también afecta al momento de entrenar al modelo, si bien escoger una métrica apropiada proporciona mejor conocimiento del real rendimiento, también se debe considerar que entrenar al modelo con datos desbalanceados sesga el entrenamiento hacia la clase mayoritaria y por consiguiente se hace menos eficiente al identificar nuevos datos perteneciente a la clase minoritaria. Por lo tanto se recurren a técnicas que permiten entrenar al modelo con datos de ambas clases de igual manera, conocidas como técnicas de balanceo de datos.

Al separar los datos entre conjunto de entrenamiento y conjunto de prueba, se tiene que el ajuste del modelo depende de esta partición, evaluando el rendimiento de manera sesgada sobre un conjunto menor a los datos disponibles. Se utiliza la técnica de validación cruzada, haciendo que se utilicen todos los datos disponibles para evaluar el rendimiento del modelo. Para esto se crean 10 subconjuntos de la base de datos original, de tal forma que la distribución de clases se mantenga y en cada uno de los 10 subconjuntos, en cada iteración se seleccionan 9 subconjuntos como conjunto de entrenamiento, mientras se reserva el último para evaluar el desempeño, finalmente se obtiene el rendimiento del promedio de las iteraciones. Esta estrategia se denomina cross validation leave-one-out.

4.7. Selección de Atributos

Uno de los objetivos específicos que propone el trabajo es reconocer los atributos de mayor relevancia para cada caso de estudio, atributos necesarios para evaluar los modelos de clasificación. Para esto se realiza una metodología de selección de atributos utilizando los 6 modelos de clasificación base (modelos no ensamblados) donde a través de cada uno se propone un conjunto propio de atributos relevante. Este procedimiento de detalla a continuación.

Se hace uso de la técnica forward variable selection, la que comienza con un conjunto qué inicialmente se encuentra vacío y evalúa la incorporación de cada uno de los atributos por medio iteraciones, haciendo ingreso el atributo que proporciona un mejor rendimiento del modelo, una vez que este hace ingreso se repite el proceso evaluando la incorporación de

los atributos restante al nuevo conjunto inicial, este proceso se detiene cuando ya no existen atributos que mejoren el rendimiento del modelo. Obteniendo finalmente un conjunto de atributos seleccionados que se construye maximizando la métrica de rendimiento.

Aunque, debido a la aleatoriedad del proceso de selección se realiza este procedimiento 20 veces [1] , obteniendo 20 conjuntos seleccionados, y por lo tanto se procede a seleccionar al conjunto con mayor número de repeticiones como el conjunto de atributos que maximiza el rendimiento del modelo en específico, este será llamado conjunto de atributos relevante.

Modelos de Linea Base

Para reconocer y validar el poder predictivo de los modelos con los conjuntos seleccionados es necesario utilizar modelos de linea base, este concepto corresponde al resultado de un modelo simple que sigue una regla básica. Para el trabajo de investigación se utilizan dos estrategias o reglas para establecer las lineas bases de rendimiento.

Estrategia Uniforme El modelo realiza predicciones uniformemente al azar, donde la probabilidad de pertenecer a alguna clase es la misma para cada clase, de esta forma la probabilidad de que un estudiante sea clasificado como desertor o de egreso oportuno, es del 50 %.

Estrategia Estratificada Este método clasifica a los estudiantes al azar, aplicando una función de probabilidad que mantiene la distribución de clases que presenta la base de datos originalmente, por ejemplo para el estudio de deserción al primer año, se tiene que solo el 5.4 % de los estudiantes analizados desertan, por lo tanto el modelo predice con un 5.4 % de probabilidad que un nuevo estudiante deserte al primer año.

A continuación se presentan las tablas de resultados para los modelos de linea base y los modelos de clasificación. De esta forma se pueden comparar los rendimientos obtenidos cuando los modelos son entrenado utilizando el conjunto de atributos relevante seleccionado.

Deserción al Primer Año

Cada uno de los modelos de la investigación superan los rendimientos bases propuestos.

Se tiene que el atributo correspondiente a la edad de ingreso se selecciona en cada uno de los 6 conjuntos relevantes, seguido por el atributo de región, siendo seleccionado en 5 de las 6 ocasiones y el atributo género encontrándose 3 veces, siendo estos 3 atributos demográficos. En menor medida se encuentra el atributo de admisión especial hallándose en 2 conjuntos, de los cuales ambos cuentan con edad de ingreso, región y género. Finalmente los atributos de provincia, puntaje PSU de lenguaje y puntaje nem aparecen solo en 1 conjunto relevante.

[1] Con esta cantidad de repeticiones es posible identificar con mayor claridad el conjunto más propenso a ser seleccionado, para todos los métodos de clasificación

MODELO	F1-SCORE	VARIABLES
M. Uniforme	0.1001 ± 0.13	–
M. Estratificado	0.0631 ± 0.031	–
Logistic Regresion	0.2496 ± 0.031	ADMISION_ESPECIAL, EDAD_INGRESO, GENERO, REGION
Decision Tree	0.2143 ± 0.032	ADMISION_ESPECIAL, EDAD_INGRESO, REGION
Naive Bayes	0.1909 ± 0.057	EDAD_INGRESO, GENERO
KN Nerighboors	0.2150 ± 0.054	EDAD_INGRESO, GENERO, PROVINCIA, REGION
Neural Net	0.2208 ± 0.031	EDAD_INGRESO, PTJE_PSU_LENGUAJE, REGION
S Vector Machine	0.2365 ± 0.032	EDAD_INGRESO, PTJE_NEM, REGION

Tabla 4.13: Rendimientos de métodos con su conjunto relevante. Deserción primer año

Deserción al Segundo Año

Cada uno de los modelos que de la investigación superan los rendimientos bases propuestos.

Los modelos de Super Vector Machine y Logistic Regresion seleccionan el mismo conjunto de atributos relevante. Al igual que para el estudio de la deserción al primer año, el atributo de la edad de ingreso es el atributo que se repite en la mayor cantidad de conjuntos, encontrándose en 5 de los 6 conjuntos seleccionados, seguido por la cantidad de ramos reprobados durante el primer año, apareciendo en 3 de los conjuntos. En menor frecuencia se encuentra el tipo de educación, la comuna y el atributo de admisión especial.

MODELO	F1-SCORE	VARIABLES
M. Uniforme	0.0834 ± 0.009	–
M. Estratificado	0.0419 ± 0.036	–
Logistic Regresion	0.2254 ± 0.044	EDAD_INGRESO, TIPO_EDUCACION
Decision Tree	0.2123 ± 0.007	EDAD_INGRESO, R_REPROB_1ER
Naive Bayes	0.2762 ± 0.048	EDAD_INGRESO
KN Nerighboors	0.2183 ± 0.077	COMUNA, R_REPROB_1ER
Neural Net	0.2316 ± 0.025	ADMISION_ESPECIAL, EDAD_INGRESO, R_REPROB_1ER
S Vector Machine	0.2241 ± 0.043	EDAD_INGRESO, TIPO_EDUCACION

Tabla 4.14: Rendimientos de métodos con su conjunto relevante. Deserción segundo año

Egreso Oportuno

Cada uno de los modelos que de la investigación superan los rendimientos bases propuestos.

Se tiene que el promedio de notas de la enseñanza media se encuentra en 5 de los 6 conjuntos de atributos seleccionado, el puntaje PSU electivo aparece en 4 conjuntos distintos, otros dos atributos con importante presencia son los correspondiente a la región y el tipo de

educación que recibe el estudiante en su etapa previa, estos tienen presencia en 3 conjuntos relevantes.

La edad de ingreso, nuevamente se encuentra como un atributo seleccionado por la técnica, sin embargo, a diferencia de los otros casos de estudio, este se repite solo en dos conjuntos y no con una importante presencia como en los casos anteriores. Continuando con los atributos con menos repeticiones se encuentra el puntaje ponderado de ingreso, la preferencia de postulación, puntaje NEM, puntaje PSU lenguaje y la provincia.

MODELO	FSCORE	VARIABLES
M. Uniforme	0.3258 ± 0.022	-
M. Estratificado	0.2260 ± 0.029	-
Logistic Regresion	0.4101 ± 0.021	'EDAD_INGRESO', 'PTJE_PSU_ELECTIVO' 'PREFERENCIA_POSTULACION', 'PROM_NOTA_ENSEÑANZA_MEDIA',
Decision Tree	0.4132 ± 0.021	'PROM_NOTA_ENSEÑANZA_MEDIA', 'REGION', 'TIPO_EDUCACION'
Naive Bayes	0.4074 ± 0.013	'PJE_PONDERADO', 'PROM_NOTA_ENSEÑANZA_MEDIA'
KN Neighbors	0.4103 ± 0.020	'PJE_PONDERADO', 'TIPO_EDUCACION' 'PROM_NOTA_ENSEÑANZA_MEDIA', 'PTJE_PSU_ELECTIVO', 'REGION'
Neural Net	0.4012 ± 0.026	'PTJE_NEM', 'PTJE_PSU_ELECTIVO', 'PTJE_PSU_LENGUAJE', 'REGION', 'TIPO_EDUCACION'
S Vector Machine	0.4097 ± 0.018	'EDAD_INGRESO', 'PROM_NOTA_ENSEÑANZA_MEDIA', 'PROVINCIA', 'PTJE_PSU_ELECTIVO'

Tabla 4.15: Rendimientos de métodos con su conjunto relevante, Egreso Oportuno

Finalmente de esta forma se obtiene un conjunto atributos para cada uno de los modelos evaluados, si bien, cada uno de estos se considera como relevante para explicar el fenómeno, no es posible priorizar uno sobre otro y por lo tanto se les considera por igual. Estos serán útiles para evaluar el rendimiento de los modelos bajo distintos atributos seleccionados.

4.8. Evaluación de Modelos de Clasificación

La investigación responde a ser útil para el desarrollo de un prototipo como herramienta de apoyo para la gestión de permanencia, debido a esto, dependiendo de los recursos e intereses se pueden priorizar distintas métricas, ya sea si se debe priorizar la identificación precisa por limitaciones presupuestaria (precision) o la identificación de la mayor cantidad de estudiantes propensos a presentar los casos estudiados (recall).

Por esta razón, los modelos son evaluados en 4 dimensiones, donde la principal métrica de rendimiento corresponde a f1-score, mientras que las métricas de accuracy, precision y recall

sirven como métricas de soporte.

Utilizando cada uno de los conjuntos relevantes se procede a obtener los rendimientos de los modelos cuando son entrenados. Las métricas de desempeño son obtenidas utilizando validación cruzada, y por lo tanto estos valores corresponden a valores medios. Con el fin de examinar si existen diferencias de rendimiento estadísticamente significativas entre los modelos se realiza el test de ANOVA en cada una de las métricas [32].

El test de ANOVA de una vía compara las medias entre los grupos y determina si alguna de estas medias es estadísticamente diferente. Con una hipótesis nula de la siguiente forma, con μ = media del grupo y k = número de grupos.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{4.3}$$

Si el test de ANOVA entrega un resultado estadísticamente significativo, entonces se rechaza la hipótesis nula, existiendo al menos dos medias de grupo que son estadísticamente diferentes entre sí.

Deserción al Primer Año

Se cuenta con seis conjuntos relevantes, los que cuentan con distintas combinaciones de atributos.

conjunto 1: 'ADMISION ESPECIAL', 'EDAD DE INGRESO', 'GENERO', 'REGION'

conjunto 2: 'ADMISION ESPECIAL', 'EDAD DE INGRESO', 'REGION'

conjunto 3: 'EDAD DE INGRESO', 'GENERO'

conjunto 4: 'EDAD DE INGRESO', 'GENERO', 'PROVINCIA', 'REGION'

conjunto 5: 'EDAD DE INGRESO', 'PTJE. PSU DE LENGUAJE', 'REGION'

conjunto 6: 'EDAD DE INGRESO', 'PTJE N.E.M.', 'REGION'

C1 :	ADMISION ESPECIAL', 'EDAD DE INGRESO', 'GENERO', 'REGION			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.2351 ± 0.032	0.8221 ± 0.039	0.1558 ± 0.024	0.4947 ± 0.086
DecisionTree	0.2084 ± 0.032	0.7809 ± 0.020	0.1301 ± 0.021	0.5274 ± 0.078
NaiveBayes	0.1595 ± 0.034	0.9288 ± 0.010	0.2478 ± 0.085	0.1242 ± 0.032
KNeighbors	0.2305 ± 0.040	0.9072 ± 0.008	0.2128 ± 0.044	0.2546 ± 0.046
NeuralNet	0.2168 ± 0.027	0.7741 ± 0.012	0.1337 ± 0.017	0.5741 ± 0.079
SVectorMachine	0.2185 ± 0.047	0.8182 ± 0.042	0.1460 ± 0.036	0.4481 ± 0.067
BaggingTree	0.2064 ± 0.035	0.7680 ± 0.028	0.1275 ± 0.023	0.5497 ± 0.085
BoostedTree	0.2073 ± 0.028	0.7699 ± 0.023	0.1280 ± 0.018	0.5477 ± 0.060
Stacking	0.1198 ± 0.015	0.3631 ± 0.133	0.0653 ± 0.010	0.7757 ± 0.106

Tabla 4.16: Resultados de rendimiento, C.E. Deserción primer año en conjunto 1

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=2.9115e-10 . Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=1.1059e-37. Se **rechaza** hipótesis nula para métrica **precision**, p-value=8.4918e-15. Se **rechaza** hipótesis nula para métrica **recall**, p-value=2.4774e-30.

C2 :	'ADMISION ESPECIAL', 'EDAD DE INGRESO', 'REGION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.2280 ± 0.033	0.8165 ± 0.042	0.1511 ± 0.025	0.4971 ± 0.126
DecisionTree	0.2171 ± 0.032	0.7855 ± 0.015	0.1357 ± 0.020	0.5457 ± 0.085
NaiveBayes	0.1921 ± 0.061	0.9253 ± 0.007	0.2385 ± 0.063	0.1691 ± 0.078
KNeighbors	0.2055 ± 0.037	0.9139 ± 0.007	0.2091 ± 0.042	0.2036 ± 0.037
NeuralNet	0.2213 ± 0.033	0.7801 ± 0.016	0.1373 ± 0.021	0.5721 ± 0.082
SVectorMachine	0.2214 ± 0.052	0.8457 ± 0.035	0.1568 ± 0.037	0.4072 ± 0.131
BaggingTree	0.2156 ± 0.030	0.7809 ± 0.018	0.1343 ± 0.020	0.5497 ± 0.066
BoostedTree	0.2136 ± 0.032	0.7811 ± 0.016	0.1330 ± 0.020	0.5435 ± 0.077
Stacking	0.1686 ± 0.053	0.5383 ± 0.230	0.0989 ± 0.037	0.7292 ± 0.161

Tabla 4.17: Resultados de rendimiento, C.E. Deserción primer año en conjunto 2

No se rechaza hipótesis nula para métrica **f1-score**, p-value=0.1058 . Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=5.5966e-15. Se **rechaza** hipótesis nula para métrica **precision**, p-value=2.9863e-12. Se **rechaza** hipótesis nula para métrica **recall**, p-value=9.1895e-21.

C3 :	'EDAD DE INGRESO', 'GENERO'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.1692 ± 0.038	0.7445 ± 0.063	0.1054 ± 0.028	0.4622 ± 0.090
DecisionTree	0.1857 ± 0.037	0.8100 ± 0.015	0.1214 ± 0.025	0.3969 ± 0.078
NaiveBayes	0.1971 ± 0.052	0.8880 ± 0.013	0.1628 ± 0.042	0.2524 ± 0.070
KNeighbors	0.1417 ± 0.047	0.6726 ± 0.127	0.0868 ± 0.034	0.4542 ± 0.137
NeuralNet	0.1686 ± 0.036	0.7492 ± 0.052	0.1045 ± 0.025	0.454 ± 0.083
SVectorMachine	0.1782 ± 0.046	0.7454 ± 0.130	0.1135 ± 0.034	0.4622 ± 0.131
BaggingTree	0.1808 ± 0.030	0.8051 ± 0.013	0.1173 ± 0.019	0.3969 ± 0.078
BoostedTree	0.1802 ± 0.031	0.7893 ± 0.023	0.1145 ± 0.019	0.4296 ± 0.104
Stacking	0.1767 ± 0.046	0.7524 ± 0.070	0.1111 ± 0.034	0.4622 ± 0.081

Tabla 4.18: Resultados de rendimiento, C.E. Deserción primer año en conjunto 3

No se rechaza hipótesis nula para métrica **f1-score**, p-value= 0.3020 . Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=2.7699e-06. Se **rechaza** hipótesis nula para métrica **precision**, p-value=0.0003. Se **rechaza** hipótesis nula para métrica **recall**, p-value=0.0002.

C4 :	'EDAD DE INGRESO', 'GENERO', 'PROVINCIA', 'REGION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.2415 ± 0.032	0.8310 ± 0.008	0.1598 ± 0.020	0.4947 ± 0.075
DecisionTree	0.2100 ± 0.033	0.7801 ± 0.019	0.1309 ± 0.021	0.5335 ± 0.078
NaiveBayes	0.1683 ± 0.072	0.9303 ± 0.008	0.2478 ± 0.100	0.1324 ± 0.061
KNeighbors	0.2111 ± 0.038	0.8991 ± 0.011	0.1859 ± 0.032	0.2486 ± 0.053
NeuralNet	0.2161 ± 0.028	0.7722 ± 0.011	0.1330 ± 0.017	0.5762 ± 0.076
SVectorMachine	0.2365 ± 0.040	0.8277 ± 0.041	0.1594 ± 0.032	0.4806 ± 0.094
BaggingTree	0.2155 ± 0.033	0.7732 ± 0.022	0.1331 ± 0.021	0.5701 ± 0.091
BoostedTree	0.2077 ± 0.030	0.7783 ± 0.026	0.1298 ± 0.021	0.5273 ± 0.061
Stacking	0.1584 ± 0.034	0.5531 ± 0.192	0.0910 ± 0.023	0.7062 ± 0.143

Tabla 4.19: Resultados de rendimiento, C.E. Deserción primer año en conjunto 4

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=0.0002 . Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=9.9531e-18. Se **rechaza** hipótesis nula para métrica **precision**, p-value=3.5205e-10. Se **rechaza** hipótesis nula para métrica **recall**, p-value=2.1131e-24

C5 :	'EDAD DE INGRESO', 'PTJE PSU LENGUAJE', 'REGION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.2446 ± 0.026	0.8452 ± 0.014	0.1671 ± 0.017	0.4623 ± 0.080
DecisionTree	0.1534 ± 0.023	0.6873 ± 0.030	0.0901 ± 0.013	0.5210 ± 0.097
NaiveBayes	0.1840 ± 0.036	0.9266 ± 0.010	0.2497 ± 0.074	0.1506 ± 0.030
KNeighbors	0.1575 ± 0.024	0.7004 ± 0.020	0.0931 ± 0.014	0.5153 ± 0.096
NeuralNet	0.2228 ± 0.033	0.7859 ± 0.017	0.1391 ± 0.021	0.562 ± 0.083
SVectorMachine	0.2299 ± 0.034	0.8409 ± 0.021	0.1570 ± 0.021	0.4419 ± 0.103
BaggingTree	0.1572 ± 0.014	0.6644 ± 0.032	0.0912 ± 0.009	0.5723 ± 0.054
BoostedTree	0.1330 ± 0.013	0.6024 ± 0.031	0.0755 ± 0.007	0.5601 ± 0.076
Stacking	0.1305 ± 0.021	0.4529 ± 0.163	0.0730 ± 0.016	0.7166 ± 0.136

Tabla 4.20: Resultados de rendimiento, C.E. Deserción primer año en conjunto 5

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=1.9296e-18. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=2.2975e-29. Se **rechaza** hipótesis nula para métrica **precision**, p-value=1.1614e-23. Se **rechaza** hipótesis nula para métrica **recall**, p-value=2.1131e-24

C6 :	'EDAD DE INGRESO', 'PTJE N.E.M.', 'REGION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.2417 ± 0.038	0.8291 ± 0.031	0.1612 ± 0.028	0.4926 ± 0.076
DecisionTree	0.2038 ± 0.026	0.7892 ± 0.038	0.1296 ± 0.019	0.4867 ± 0.052
NaiveBayes	0.1914 ± 0.041	0.9242 ± 0.008	0.2358 ± 0.063	0.1649 ± 0.037
KNeighbors	0.2053 ± 0.018	0.8211 ± 0.015	0.1357 ± 0.009	0.4278 ± 0.078
NeuralNet	0.2199 ± 0.031	0.7774 ± 0.013	0.1360 ± 0.019	0.5761 ± 0.088
SVectorMachine	0.2267 ± 0.036	0.8454 ± 0.033	0.1597 ± 0.027	0.4179 ± 0.111
BaggingTree	0.1800 ± 0.015	0.7382 ± 0.022	0.1088 ± 0.010	0.5254 ± 0.042
BoostedTree	0.1721 ± 0.017	0.7357 ± 0.025	0.1039 ± 0.010	0.5049 ± 0.074
Stacking	0.1580 ± 0.033	0.5492 ± 0.186	0.0906 ± 0.023	0.7105 ± 0.133

Tabla 4.21: Resultados de rendimiento, C.E. Deserción primer año en conjunto 6

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=4.1221e-07 . Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=2.8854e-17. Se **rechaza** hipótesis nula para métrica **precision**, p-value=8.2174e-17. Se **rechaza** hipótesis nula para métrica **recall**, p-value=2.3743e-20.

Se tiene que al aplicar el test de ANOVA para comparar las medias de las métricas accuracy, precision y recall, de esta manera se rechaza la hipótesis de igualdad en cada uno de los conjuntos, mientras para la métrica f1-score solo en dos grupos de atributos no es posible rechazar la hipótesis nula, para el conjunto 2 y 3.

Con respecto a la métrica f1-score, para el quinto conjunto de atributos los modelos Super Vector Machine, Logistic Regresion y Redes Neuronales presentan un rendimiento superior, mientras qué para el primer conjunto se tiene que el modelo de Stacking Generalizer obtiene rendimientos inferiores a los demás.

Para a la métrica accuracy, se tiene que en cada uno de los conjuntos, el modelo de Naive Bayes presenta un rendimiento superior, además se distingue el buen rendimiento del modelo de KN Neighbors en los conjuntos 1,2 y 4. Por otra, parte el modelo Stacking Generalizer presenta rendimientos inferiores en todos los conjuntos menos en el conjunto 3, el que solo cuenta con los atributos edad de ingreso y género.

Al analizar los rendimientos de la métrica precision, se tiene que el modelo Naive Bayes presenta un rendimiento destacable en cada uno de los conjuntos de atributos.

Los rendimientos en recall del modelo de Stacking Generalizer presenta considerables valores, mientras que el modelo Naive Bayes presenta rendimientos inferiores, esto ocurre en todos los conjuntos menos en el tercero, el que cuenta con menos información al contar con solo dos atributos.

Deserción al Segundo Año

Se cuenta con cinco conjuntos relevantes distintos, el conjunto seleccionado por el modelo de Regresión Logística y Super Vector Machine coinciden, y por lo tanto solo se permite uno

de estos para evaluar y comparar los rendimientos.

conjunto 1: 'EDAD DE INGRESO', 'TIPO DE EDUCACION'

conjunto 2: 'EDAD DE INGRESO', 'RAMOS REPROB 1ER AÑO'

conjunto 3: 'EDAD DE INGRESO'

conjunto 4: 'COMUNA', 'RAMOS REPROB. 1ER AÑO'

conjunto 5: 'ADMISION ESPECIAL', 'EDAD DE INGRESO', 'RAMOS REPROB 1ER AÑO'

C1 :	'EDAD DE INGRESO', 'TIPO DE EDUCACION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.2212 ± 0.047	0.8460 ± 0.021	0.1445 ± 0.032	0.4755 ± 0.097
DecisionTree	0.2174 ± 0.047	0.8452 ± 0.022	0.1422 ± 0.033	0.4641 ± 0.085
NaiveBayes	0.2616 ± 0.046	0.9043 ± 0.020	0.2066 ± 0.039	0.3704 ± 0.084
KNeighbors	0.1261 ± 0.060	0.5863 ± 0.151	0.0735 ± 0.043	0.5735 ± 0.140
NeuralNet	0.2177 ± 0.045	0.8403 ± 0.022	0.1411 ± 0.030	0.4810 ± 0.088
SVectorMachine	0.2207 ± 0.047	0.8475 ± 0.020	0.1446 ± 0.031	0.4698 ± 0.096
BaggingTree	0.2168 ± 0.038	0.8440 ± 0.014	0.1410 ± 0.025	0.4726 ± 0.089
BoostedTree	0.2108 ± 0.035	0.8393 ± 0.013	0.1360 ± 0.022	0.4726 ± 0.091
Stacking	0.2222 ± 0.042	0.8501 ± 0.014	0.1458 ± 0.028	0.4698 ± 0.096

Tabla 4.22: Resultados de rendimiento, C.E. Deserción segundo año en conjunto 1

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=1.9724e-05. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=2.5993e-19. Se **rechaza** hipótesis nula para métrica **precision**, p-value=3.1003e-09. Se **rechaza** hipótesis nula para métrica **recall**, p-value=0.0204.

C2 :	['EDAD DE INGRESO', 'RAMOS REPROB. 1ER AÑO']			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.1983 ± 0.011	0.7642 ± 0.015	0.1174 ± 0.006	0.6409 ± 0.066
DecisionTree	0.2303 ± 0.026	0.8167 ± 0.027	0.1433 ± 0.019	0.5960 ± 0.073
NaiveBayes	0.2497 ± 0.031	0.8438 ± 0.006	0.1598 ± 0.019	0.5732 ± 0.093
KNeighbors	0.1272 ± 0.055	0.5435 ± 0.155	0.0735 ± 0.039	0.6300 ± 0.091
NeuralNet	0.2009 ± 0.013	0.7643 ± 0.015	0.1189 ± 0.008	0.6495 ± 0.056
SVectorMachine	0.2089 ± 0.022	0.7821 ± 0.032	0.1261 ± 0.016	0.6243 ± 0.070
BaggingTree	0.2042 ± 0.023	0.7866 ± 0.023	0.1233 ± 0.014	0.6020 ± 0.104
BoostedTree	0.2262 ± 0.034	0.8066 ± 0.029	0.1393 ± 0.023	0.6130 ± 0.088
Stacking	0.2389 ± 0.025	0.8196 ± 0.022	0.1484 ± 0.016	0.6184 ± 0.077

Tabla 4.23: Resultados de rendimiento, C.E. Deserción segundo año en conjunto 2

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=7.5008e-12. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=9.3227e-18. Se **rechaza** hipótesis nula para

métrica **precision**, p-value=3.0146e-12. **No se rechaza** hipótesis nula para métrica **recall**, p-value=0.6439.

C3 :	'EDAD DE INGRESO'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.2090 ± 0.037	0.8306 ± 0.014	0.1330 ± 0.023	0.4896 ± 0.086
DecisionTree	0.2064 ± 0.038	0.8309 ± 0.013	0.1314 ± 0.024	0.4811 ± 0.089
NaiveBayes	0.2761 ± 0.048	0.9148 ± 0.006	0.2258 ± 0.036	0.3594 ± 0.078
KNeighbors	0.1420 ± 0.071	0.5581 ± 0.232	0.0885 ± 0.058	0.6111 ± 0.166
NeuralNet	0.2090 ± 0.037	0.8306 ± 0.014	0.1330 ± 0.023	0.4896 ± 0.086
SVectorMachine	0.2090 ± 0.037	0.8306 ± 0.014	0.1330 ± 0.023	0.4896 ± 0.086
BaggingTree	0.2153 ± 0.054	0.8385 ± 0.031	0.1428 ± 0.049	0.4668 ± 0.071
BoostedTree	0.2067 ± 0.036	0.8295 ± 0.013	0.1313 ± 0.023	0.4867 ± 0.088
Stacking	0.2084 ± 0.037	0.8300 ± 0.014	0.1325 ± 0.023	0.4896 ± 0.086

Tabla 4.24: Resultados de rendimiento, C.E. Deserción segundo año en conjunto 3

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=6.1707e-05. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=1.3931e-12. Se **rechaza** hipótesis nula para métrica **precision**, p-value=1.9924e-09. Se **rechaza** hipótesis nula para métrica **recall**, p-value=0.0007.

C4 :	'COMUNA', 'RAMOS REPROB. 1ER AÑO'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.1890 ± 0.023	0.7668 ± 0.022	0.1126 ± 0.014	0.5930 ± 0.065
DecisionTree	0.1760 ± 0.024	0.7906 ± 0.029	0.1080 ± 0.016	0.4853 ± 0.064
NaiveBayes	0.1438 ± 0.030	0.6815 ± 0.060	0.0827 ± 0.018	0.5761 ± 0.127
KNeighbors	0.2172 ± 0.084	0.9043 ± 0.040	0.1879 ± 0.080	0.2796 ± 0.107
NeuralNet	0.1803 ± 0.027	0.7578 ± 0.033	0.1071 ± 0.017	0.5788 ± 0.084
SVectorMachine	0.1562 ± 0.020	0.7096 ± 0.036	0.0903 ± 0.011	0.5871 ± 0.091
BaggingTree	0.1633 ± 0.028	0.7381 ± 0.058	0.0971 ± 0.019	0.5478 ± 0.104
BoostedTree	0.1903 ± 0.032	0.7969 ± 0.028	0.1170 ± 0.021	0.5168 ± 0.078
Stacking	0.0678 ± 0.020	0.2922 ± 0.066	0.0361 ± 0.010	0.5646 ± 0.176

Tabla 4.25: Resultados de rendimiento, C.E. Deserción segundo año en conjunto 4

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=1.0890e-10. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=2.9179e-43. Se **rechaza** hipótesis nula para métrica **precision**, p-value=5.2605e-13. Se **rechaza** hipótesis nula para métrica **recall**, p-value=8.4136e-08.

C5 :	'ADMISION ESPECIAL', 'EDAD DE INGRESO' 'RAMOS REPROB 1ER AÑO'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.1993 ± 0.015	0.7682 ± 0.016	0.1184 ± 0.009	0.6326 ± 0.063
DecisionTree	0.2316 ± 0.021	0.8206 ± 0.015	0.1443 ± 0.013	0.5934 ± 0.076
NaiveBayes	0.2421 ± 0.033	0.8301 ± 0.016	0.1524 ± 0.022	0.5929 ± 0.084
KNeighbors	0.1628 ± 0.065	0.6539 ± 0.189	0.0986 ± 0.047	0.5990 ± 0.101
NeuralNet	0.2023 ± 0.013	0.7656 ± 0.021	0.1200 ± 0.009	0.6496 ± 0.051
SVectorMachine	0.2006 ± 0.014	0.7726 ± 0.022	0.1197 ± 0.008	0.6268 ± 0.081
BaggingTree	0.2187 ± 0.014	0.8001 ± 0.020	0.1335 ± 0.010	0.6130 ± 0.066
BoostedTree	0.2161 ± 0.027	0.7991 ± 0.031	0.1323 ± 0.018	0.5990 ± 0.063
Stacking	0.2190 ± 0.025	0.7918 ± 0.027	0.1328 ± 0.017	0.6328 ± 0.046

Tabla 4.26: Resultados de rendimiento, C.E. Deserción segundo año en conjunto 5

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=2.5580e-05. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=1.7738e-05. Se **rechaza** hipótesis nula para métrica **precision**, p-value=3.6366e-05. **No se rechaza** hipótesis nula para métrica **recall**, p-value=0.6585.

Con respecto a la métrica f1-score, se halla que el modelo de KN Neighbors presenta un rendimiento inferior a los demás modelos, esto se repite en todos los conjuntos menos en el tercero, el que solo cuenta con la edad de ingreso.

Para la métrica de accuracy, el modelo de Naive Bayes obtiene buenos rendimientos para 3 de los 5 conjuntos, mientras que nuevamente se tiene el modelo de KN Neighbors muestras rendimientos inferiores, esto sucede en cada uno de los conjuntos menos en el cuarto, donde ocurre lo contrario, se muestra KN Neighbors con buenos resultados, mientras que Stacking se muestra como un modelo de mal rendimiento (el cuarto conjunto solo cuenta con las variables de comuna y la cantidad de ramos reprobados en el primer año).

Al comparar los rendimientos obtenidos en la métrica precision se encuentra que el modelo de Naive Bayes muestra rendimientos superiores, mientras que ocurre lo opuesto para el modelo de KN Neighbors, esto se repite en 4 de los 5 conjuntos. Para el conjunto número 4 (comuna y ramos reprobados) el modelo de KN Neighbors se sobrepone a los otros modelos mientras que Stacking Generalizer es el cual muestra peor rendimiento.

Referente a la métrica de recall, se tiene 2 conjuntos de atributos donde no es posible rechazar la hipótesis nula. Por otra parte, en los conjuntos 1 y 2 se tiene claramente que el modelo de KN Neighbors muestra mejores resultados que el modelo Naive Bayes. Sin embargo, para el conjunto 4 el modelo de KN Neighbors obtiene el peor rendimiento que todos los demás modelos estudiados.

Egreso Oportuno

Se cuenta con seis conjuntos relevantes, los que cuentan con distintas combinaciones de atributos.

conjunto 1: 'EDAD DE INGRESO', 'PREFERENCIA DE POSTULACION', 'PROM. NOTA ESEÑANZA MEDIA', 'PTJE. PSU ELECTIVO'

conjunto 2: 'PROM. NOTA ESEÑANZA MEDIA', 'REGION', 'TIPO DE EDUCACION'

conjunto 3: 'PTJE. PONDERADO', 'PROM. NOTA ESEÑANZA MEDIA'

conjunto 4: 'PTJE. PONDERADO', 'PROM. NOTA ESEÑANZA MEDIA', 'PTJE PSU ELECTIVO', 'REGION', 'TIPO DE EDUCACION'

conjunto 5: 'PTJE. N.E.M.', 'PTJE. PSU ELECTIVO', 'PTJE. PSU LENGUAJE', 'REGION', 'TIPO DE EDUCACION'

conjunto 6: 'EDAD DE INGRESO', 'PROM. NOTA ENSEÑANZA MEDIA', 'PROVINCIA', 'PTJE. PSU ELECTIVO'

C1 :	'EDAD DE INGRESO', 'PREFERENCIA DE POSTULACION', 'PROM. NOTA ENSEÑANZA MEDIA', 'PTJE PSU ELECTIVO'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.4109 ± 0.017	0.5237 ± 0.017	0.2910 ± 0.012	0.6990 ± 0.036
DecisionTree	0.3706 ± 0.024	0.5441 ± 0.023	0.2760 ± 0.018	0.5648 ± 0.043
NaiveBayes	0.4013 ± 0.016	0.4070 ± 0.020	0.2641 ± 0.010	0.8367 ± 0.045
KNeighbors	0.3773 ± 0.015	0.5386 ± 0.017	0.2779 ± 0.012	0.5880 ± 0.023
NeuralNet	0.4124 ± 0.019	0.5170 ± 0.019	0.2901 ± 0.013	0.7136 ± 0.041
SVectorMachine	0.4100 ± 0.014	0.5214 ± 0.011	0.2899 ± 0.010	0.6999 ± 0.028
BaggingTree	0.3878 ± 0.026	0.5342 ± 0.021	0.2820 ± 0.019	0.6211 ± 0.046
BoostedTree	0.3836 ± 0.028	0.5258 ± 0.017	0.2774 ± 0.019	0.6218 ± 0.053
Stacking	0.4074 ± 0.012	0.4873 ± 0.015	0.2809 ± 0.008	0.7417 ± 0.034

Tabla 4.27: Resultados de rendimiento, C.E. Egreso oportuno en conjunto 1

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=7.9367e-07. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=4.0629e-21. Se **rechaza** hipótesis nula para métrica **precision**, p-value=0.0039. Se **rechaza** hipótesis nula para métrica **recall**, p-value=6.2653e-22.

C2 :	'PROM. NOTA ENSEÑANZA MEDIA', 'REGION', 'TIPO DE EDUCACION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.4028 ± 0.020	0.4988 ± 0.023	0.2811 ± 0.014	0.7111 ± 0.041
DecisionTree	0.4152 ± 0.015	0.4774 ± 0.018	0.2829 ± 0.010	0.7811 ± 0.042
NaiveBayes	0.3884 ± 0.007	0.2699 ± 0.027	0.2426 ± 0.005	0.9750 ± 0.016
KNeighbors	0.3710 ± 0.031	0.5514 ± 0.036	0.2789 ± 0.023	0.5582 ± 0.072
NeuralNet	0.4100 ± 0.017	0.4858 ± 0.028	0.2822 ± 0.012	0.7521 ± 0.051
SVectorMachine	0.4021 ± 0.012	0.4965 ± 0.021	0.2803 ± 0.006	0.7134 ± 0.051
BaggingTree	0.4118 ± 0.021	0.4795 ± 0.028	0.2817 ± 0.015	0.7666 ± 0.052
BoostedTree	0.4180 ± 0.022	0.4770 ± 0.018	0.2841 ± 0.014	0.7908 ± 0.049
Stacking	0.4101 ± 0.015	0.4965 ± 0.042	0.2852 ± 0.014	0.7360 ± 0.062

Tabla 4.28: Resultados de rendimiento, C.E. Egreso oportuno en conjunto 2

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=1.9707e-05. Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=4.5680e-33. Se **rechaza** hipótesis nula para métrica **precision**, p-value=4.3123e-08. Se **rechaza** hipótesis nula para métrica **recall**, p-value=4.1912e-25.

C3 :	'PJE. PONDERADO', 'PROM. NOTA ENSEÑANZA MEDIA'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.3986 ± 0.017	0.5282 ± 0.020	0.2862 ± 0.011	0.6588 ± 0.050
DecisionTree	0.3784 ± 0.022	0.5267 ± 0.019	0.2752 ± 0.016	0.6065 ± 0.041
NaiveBayes	0.4070 ± 0.020	0.5049 ± 0.024	0.2847 ± 0.016	0.7144 ± 0.034
KNeighbors	0.3884 ± 0.013	0.5372 ± 0.016	0.2833 ± 0.009	0.6186 ± 0.032
NeuralNet	0.4016 ± 0.025	0.5193 ± 0.021	0.2852 ± 0.016	0.6797 ± 0.056
SVectorMachine	0.3989 ± 0.023	0.5170 ± 0.023	0.2834 ± 0.017	0.6741 ± 0.039
BaggingTree	0.3870 ± 0.026	0.5326 ± 0.019	0.2811 ± 0.018	0.6210 ± 0.047
BoostedTree	0.3708 ± 0.027	0.5219 ± 0.017	0.2697 ± 0.018	0.5937 ± 0.055
Stacking	0.4045 ± 0.015	0.4994 ± 0.015	0.2820 ± 0.010	0.7160 ± 0.040

Tabla 4.29: Resultados de rendimiento, C.E. Egreso oportuno en conjunto 3

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=0.0066 Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=0.0019. **No se rechaza** hipótesis nula para métrica **precision**, p-value=0.3689. Se **rechaza** hipótesis nula para métrica **recall**, p-value=4.6213e-09.

C4 :	'PJE PONDERADO', 'PROM. NOTA ENSEÑANZA MEDIA', 'PTJE. PSU ELECTIVO', 'REGION', 'TIPO DE EDUCACION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.4010 ± 0.020	0.5390 ± 0.025	0.2904 ± 0.017	0.6484 ± 0.025
DecisionTree	0.3963 ± 0.033	0.5330 ± 0.023	0.2860 ± 0.022	0.6467 ± 0.069
NaiveBayes	0.3118 ± 0.144	0.3506 ± 0.202	0.2450 ± 0.059	0.7903 ± 0.389
KNeighbors	0.4102 ± 0.026	0.5554 ± 0.022	0.2996 ± 0.019	0.6508 ± 0.049
NeuralNet	0.4049 ± 0.030	0.5252 ± 0.022	0.2883 ± 0.020	0.6805 ± 0.058
SVectorMachine	0.4043 ± 0.019	0.5279 ± 0.023	0.2890 ± 0.014	0.6741 ± 0.039
BaggingTree	0.4161 ± 0.021	0.5680 ± 0.021	0.3067 ± 0.016	0.6476 ± 0.037
BoostedTree	0.4161 ± 0.028	0.5590 ± 0.018	0.3035 ± 0.019	0.6621 ± 0.053
Stacking	0.4028 ± 0.021	0.5110 ± 0.020	0.2838 ± 0.013	0.6950 ± 0.057

Tabla 4.30: Resultados de rendimiento, C.E. Egreso oportuno en conjunto 4

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=0.0031 Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=1.1864e-07. Se **rechaza** hipótesis nula para métrica **precision**, p-value=0.0002. **No se rechaza** hipótesis nula para métrica **recall**, p-value=0.4566.

C5 :	'PTJE N.E.M.', 'PTJE PSU ELECTIVO', 'PTJE PSU LENGUAJE', 'REGION', 'TIPO DE EDUCACION'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.3945 ± 0.015	0.5269 ± 0.014	0.2835 ± 0.009	0.6491 ± 0.041
DecisionTree	0.3917 ± 0.014	0.5390 ± 0.030	0.2860 ± 0.013	0.6243 ± 0.037
NaiveBayes	0.3866 ± 0.007	0.2682 ± 0.034	0.2415 ± 0.006	0.9694 ± 0.027
KNeighbors	0.4043 ± 0.022	0.5470 ± 0.017	0.2941 ± 0.014	0.6476 ± 0.046
NeuralNet	0.4081 ± 0.019	0.5326 ± 0.017	0.2920 ± 0.014	0.6781 ± 0.040
SVectorMachine	0.4060 ± 0.020	0.5258 ± 0.013	0.2890 ± 0.012	0.6830 ± 0.049
BaggingTree	0.4032 ± 0.027	0.5621 ± 0.017	0.2981 ± 0.018	0.6234 ± 0.052
BoostedTree	0.3950 ± 0.034	0.5309 ± 0.029	0.2849 ± 0.025	0.6444 ± 0.061
Stacking	0.4067 ± 0.034	0.5145 ± 0.019	0.2863 ± 0.022	0.7024 ± 0.075

Tabla 4.31: Resultados de rendimiento, C.E. Egreso oportuno en conjunto 5

No se rechaza hipótesis nula para métrica **f1-score**, p-value=0.4587 Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=5.4685e-44. **No se rechaza** hipótesis nula para métrica **precision**, p-value=3.8505e-09. Se **rechaza** hipótesis nula para métrica **recall**, p-value=1.2854e-25.

C6 :	'EDAD DE INGRESO', 'PROM NOTA ENSEÑANZA MEDIA', 'PROVINCIA', 'PTJE PSU ELECTIVO'			
MODELO	F1 SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	0.4109 ± 0.023	0.5246 ± 0.025	0.2914 ± 0.017	0.6975 ± 0.044
DecisionTree	0.3673 ± 0.029	0.5351 ± 0.018	0.2714 ± 0.019	0.5688 ± 0.057
NaiveBayes	0.4067 ± 0.014	0.4231 ± 0.017	0.2691 ± 0.009	0.8318 ± 0.029
KNeighbors	0.3656 ± 0.018	0.5279 ± 0.019	0.2687 ± 0.014	0.5727 ± 0.040
NeuralNet	0.4084 ± 0.014	0.5177 ± 0.020	0.2884 ± 0.009	0.7007 ± 0.042
SVectorMachine	0.4065 ± 0.024	0.5160 ± 0.019	0.2868 ± 0.016	0.6983 ± 0.052
BaggingTree	0.3757 ± 0.028	0.5244 ± 0.021	0.2730 ± 0.020	0.6026 ± 0.051
BoostedTree	0.3890 ± 0.022	0.5330 ± 0.014	0.2822 ± 0.014	0.6266 ± 0.050
Stacking	0.4088 ± 0.013	0.4938 ± 0.013	0.2830 ± 0.008	0.7369 ± 0.038

Tabla 4.32: Resultados de rendimiento, C.E. Egreso oportuno en conjunto 6

Se **rechaza** hipótesis nula para métrica **f1-score**, p-value=7.9367e-07 Se **rechaza** hipótesis nula para métrica **accuracy**, p-value=4.0629e-21. Se **rechaza** hipótesis nula para métrica **precision**, p-value=0.0039. Se **rechaza** hipótesis nula para métrica **recall**, p-value=6.2653e-22.

Si bien en general, al realizar el test de ANOVA sobre las métricas de rendimiento la hipótesis nula se rechaza, en 3 ocasiones ocurre que al realizar el test de hipótesis este no se puede rechazar, esto ocurre para la métrica precision, recall y f1-score en el tercer, cuarto y quinto conjunto respectivamente.

Al evaluar el rendimiento de la métrica f1-score, se nota que ninguno de los modelos pre-

senta un rendimiento notablemente superior, por otra parte sí se logra diferenciar modelos con rendimientos inferiores, aquí se encuentra KN Neighbors mostrando un rendimiento regular para los conjuntos 1 y 2 y claramente inferior para el sexto, Decision Tree muestra rendimientos inferiores en el conjunto 1 y 6, y por último se tiene Naive Bayes muestra rendimientos inferiores dentro del cuarto conjunto de atributos.

Con respecto a la métrica accuracy, dentro del segundo conjunto se aprecia un notable rendimiento de un modelo sobre los demás, este corresponde KN Neighbors. Por otra parte, nuevamente se tiene al algoritmo de Naive Bayes exhibiendo rendimientos inferiores, esto ocurre en 5 de los 6 conjuntos, a excepción del tercero, correspondiente a los atributos de puntaje de ingreso ponderado y promedio de enseñanza media.

Para la métrica de precision los modelos no difieren notablemente en su rendimiento, a excepción del modelo de Naive Bayes, el cual se muestra con rendimientos notablemente inferiores en los conjuntos 2, 4 y 5.

A diferencias de los resultados anteriores, el modelo de Naive Bayes muestra buenos resultados para la métrica de recall, donde este obtiene rendimientos superiores en todos los conjuntos, menos en el cuarto. Otros modelos que también destacan por su buen resultado, pero en menor medida, son los de Logistic Regresion, Neural Net, Super Vector Machine y Stacking Generalizer al utilizar los conjuntos de atributos 1 y 6. Con rendimientos inferiores se tiene al modelo Decision Tree y KN Neighbors en 3 de los 6 conjuntos que se evalúan, también se destaca el mal rendimiento en recall de los modelos de ensamblado Bagging Tree y Boosted Tree.

Capítulo 5

Comparación de Rendimientos

Como lo muestran los resultados obtenidos en los distintos test de hipótesis ANOVA, en general, existen diferencias de rendimiento en cada una de las dimensiones de desempeño entre los modelos de clasificación. Si bien, a través de los valores obtenidos es posible destacar el rendimiento de algunos modelos, se hace difícil la comparación entre estos.

Por lo tanto, con el fin realizar una comparación general se procede a generar un ranking según rendimiento obtenido para cada uno de los conjuntos relevantes identificados, considerando de manera independiente cada una de las métricas de desempeño. Al contar con varios set de entrenamiento se busca obtener un rendimiento general del modelo utilizando distintas combinaciones de atributos, y de esta forma obtener una posición promedio del modelo frente a los demás.

Se evalúan los resultados sobre 4 dimensiones, correspondientes a las métricas de desempeño utilizadas a lo largo del trabajo. A continuación se aclara la información que proporciona cada una de estas métricas con el fin de comprender de mejor manera los resultados obtenidos.

Accuarcy: Mide el porcentaje de estudiantes clasificados correctamente, considerando estudiantes de clase positiva y clase negativa.

Precision: Mide la calidad del modelo en la tarea de clasificación, corresponde al porcentaje de estudiantes bien clasificados sobre todos los que son identificados como desertores o de egreso oportuno.

Recall: Mide la capacidad de poder identificar a los estudiantes de clase positiva, correspondiendo al porcentaje de estudiantes bien identificados del total de estudiantes que realmente corresponden a desertores o de egreso oportuno.

F1-Score: Se utiliza como la combinación de las ultimas dos métricas en un solo valor, permitiendo evaluar el desempeño del modelo de una manera equitativa.

5.1. Caso de Estudio de Deserción al Primer Año

A continuación se presenta la tabla de posiciones obtenidas al ser entrenados con los conjuntos relevantes para estudiar el fenómeno de deserción al primer año. Estos conjuntos se componen de la siguiente forma.

C1: 'ADMISION ESPECIAL', 'EDAD DE INGRESO', 'GENERO', 'REGION'

C2: 'ADMISION ESPECIAL', 'EDAD DE INGRESO', 'REGION'

C3: 'EDAD DE INGRESO', 'GENERO'

C4: 'EDAD DE INGRESO', 'GENERO', 'PROVINCIA', 'REGION'

C5: 'EDAD DE INGRESO', 'PTJE. PSU DE LENGUAJE', 'REGION'

C6: 'EDAD DE INGRESO', 'PTJE N.E.M.', 'REGION'

DESERCION 1ER AÑO	F1-SCORE						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	1°	1°	7°	1°	1°	1°	2.0
Decision Tree	5°	4°	2°	6°	7°	5°	4.8
Naive Bayes	8°	8°	1°	8°	4°	6°	5.8
KN Neighbors	2°	7°	9°	5°	5°	4°	5.3
Neural Net	4°	3°	8°	3°	3°	3°	4.0
S Vector Machine	3°	2°	5°	2°	2°	2°	2.7
Bagging Tree	7°	5°	3°	4°	6°	7°	5.3
Boosted Tree	6°	6°	4°	7°	8°	8°	6.5
Stacking	9°	9°	6°	9°	9°	9°	8.5

Tabla 5.1: Posición según f1-score obtenido. Estudio de deserción al primer año

F1-SCORE: Se tiene un buen posicionamiento del modelo de Logistic Regresion sobre los demás, ubicándose siempre por sobre el resto, menos en el tercer conjunto, sin embargo este conjunto posee menor numero de atributos y menor información. También se destacan los rendimientos de los modelos Super Vector Machine y Neural Net, ubicándose de manera constante en la parte alta de la posiciones, menos en el tercer conjunto.

El modelo de Stacking Generalizer se muestra en la ultima posición, menos en el tercer conjunto.

DESERCION 1ER AÑO	ACCURACY						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	3°	4°	8°	3°	2°	3°	3.8
DecisionTree	5°	5°	2°	5°	6°	5°	4.7
NaiveBayes	1°	1°	1°	1°	1°	1°	1.0
KNeighbors	2°	2°	9°	2°	5°	4°	4.0
NeuralNet	6°	8°	6°	8°	4°	6°	6.3
SVectorMachine	4°	3°	7°	4°	3°	2°	3.8
BaggingTree	8°	7°	3°	7°	7°	7°	6.5
BoostedTree	7°	6°	4°	6°	8°	8°	6.5
Stacking	9°	9°	5°	9°	9°	9°	8.3

Tabla 5.2: Posición según accuracy obtenido. Estudio de deserción al primer año

ACCURACY: Se destaca el buen rendimiento del modelo Naive Bayes, ubicandose en cada oporunidad como el de mejor rendimiento. También se aprecia un buen rendimiento de los modelos de Logistic Regresion y KN Neighbors.

El modelo de Stacking Generalizer se muestra en la ultima posición, menos en el tercer conjunto.

DESERCION 1ER AÑO	PRECISION						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	3°	4°	7°	3°	2°	2°	3.5
DecisionTree	6°	6°	2°	7°	7°	6°	5.7
NaiveBayes	1°	1°	1°	1°	1°	1°	1.0
KNeighbors	2°	2°	9°	2°	5°	5°	4.2
NeuralNet	5°	5°	8°	6°	4°	4°	5.3
SVectorMachine	4°	3°	5°	4°	3°	3°	3.7
BaggingTree	8°	7°	3°	5°	6°	7°	6.0
BoostedTree	7°	8°	4°	8°	8°	8°	7.2
Stacking	9°	9°	6°	9°	9°	9°	8.5

Tabla 5.3: Posición según precision obtenido. Estudio de deserción al 1er año

Nuevamente se presenta el modelo de Naive Bayes ubicándose en la mejor posición en cada ocasión, también se destaca al modelo de Logistic Regresion el cual se ubica en una posición media-alta en todos los conjuntos menos en el tercero.

Los modelos Decision Tree y Stacking Generalizer se ubican reiteradamente en las posiciones inferiores.

DESERCION 1ER AÑO	RECALL						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	6°	6°	2°	6°	7°	5°	5.3
DecisionTree	5°	4°	7°	4°	5°	6°	5.2
NaiveBayes	9°	9°	9°	9°	9°	9°	9.0
KNeighbors	8°	8°	4°	8°	6°	7°	6.8
NeuralNet	2°	2°	5°	2°	3°	2°	2.7
SVectorMachine	7°	7°	3°	7°	8°	8°	6.7
BaggingTree	3°	3°	8°	3°	2°	3°	3.7
BoostedTree	4°	5°	6°	5°	4°	4°	4.7
Stacking	1°	1°	1°	1°	1°	1°	1.0

Tabla 5.4: Posición según recall obtenido. Estudio de deserción al 1er año

Con respecto a la métrica de recall, el modelo Stacking Generalizer se ubica constantemente en la mejor posición, también destaca el rendimiento del modelo Neural Net el cual se encuentra entre el segundo y tercer mejor modelo, menos para el tercer conjunto donde se ubica en una posición media.

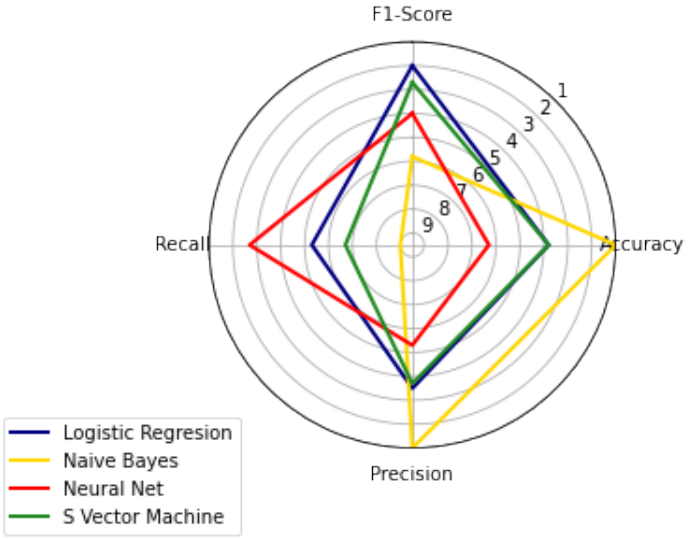
A continuación se muestra las posiciones promedio obtenidas por los modelo de clasificación bajo las diferentes dimensiones de evaluación. La métrica más importante para definir un buen rendimiento del modelo es la métrica f1-score, las restantes dan un soporte al análisis.

	POSICIÓN PROMEDIO			
MODELO	F1SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	2.0	3.8	3.5	5.3
DecisionTree	4.8	4.7	5.7	5.2
NaiveBayes	5.8	1.0	1.0	9.0
KNeighbors	5.3	4.0	4.2	6.8
NeuralNet	4.0	6.3	5.3	2.7
SVectorMachine	2.7	3.8	3.7	6.7
BaggingTree	5.3	6.5	6.0	3.7
BoostedTree	6.5	6.5	7.2	4.7
Stacking	8.5	8.3	8.5	1.0

Tabla 5.5: Resumen de posición según métrica de desempeño, deserción primer año

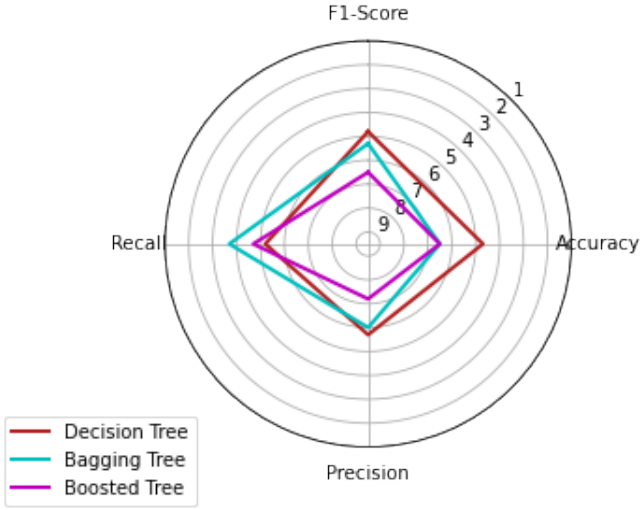
Para poder visualizar de mejor manera la comparación se procede a graficar las posiciones por medio de un gráfico radial. El primer gráfico corresponde solo a los modelos que muestran un rendimiento destacable sobre los demás, el segundo gráfico corresponde a la comparación de los modelos de Decision Tree, Baggging Tree y Boosted Tree, esto debido que los dos últimos corresponden a métodos de ensamblados que utilizan como base al método de Decision Tree. Por ultimo, el tercer gráfico permite comparar al modelo ensamblado de Stacking Generalizer en conjunto a sus modelos utilizados como base.

Figura 5.1: Resultados destacados para el estudio de deserción al primer año



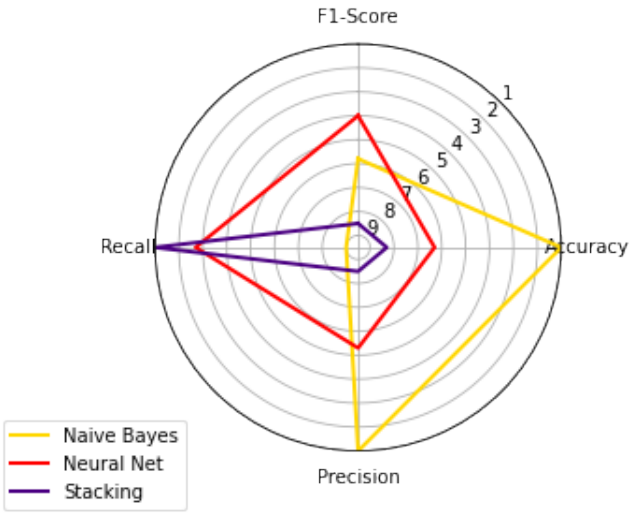
Los modelos de Logistic Regresion y Super Vector Machine obtienen los mejores rendimientos en f1-score, además cuentan con resultados similares en las métricas de soporte, también se destaca el rendimiento del modelo Neural Net, si bien este cuenta con un rendimiento medio en f1-score, destaca por su rendimiento en recall. Además, se distingue el modelo de Naive Bayes por su excelente rendimiento en las métricas accuracy y precision.

Figura 5.2: Decision Tree y metodos de ensamblaje para el estudio de deserción al primer año



De los resultados obtenidos para los métodos de ensamblado basado en Decision Tree, se observa que ninguno de estos se ubica considerablemente mejor que el modelo base, incluso el modelo Boosted Tree obtiene peores resultados.

Figura 5.3: Comparación de modelo Stacking Generalizer para el estudio de deserción al primer año



Se busca con la implementación del método de Stacking Generalizer poder combinar los poderes de predictivos de los modelos utilizados como base, claramente no se logra conseguir el resultado deseado, donde este se ubica con rendimientos inferiores.

5.2. Caso de Estudio de Deserción al Segundo Año

A continuación se presenta la tabla de posiciones sobre los diferentes conjuntos relevantes identificados al estudiar el fenómeno de deserción al segundo año. Estos conjuntos se componen de la siguiente forma.

C1: 'EDAD DE INGRESO', 'TIPO DE EDUCACION'

C2: 'EDAD DE INGRESO', 'RAMOS REPROB 1ER AÑO'

C3: 'EDAD DE INGRESO'

C4: 'COMUNA', 'RAMOS REPROB 1ER AÑO'

C5: 'ADMISION ESPECIAL', 'EDAD DE INGRESO', 'RAMOS REPROB 1ER AÑO'

DESERCION 2DO AÑO	F1-SCORE					
MODELO	C1	C2	C3	C4	C5	POSICIÓN PROMEDIO
Logistic Regresion	3°	8°	3°	3°	8°	5.0
DecisionTree	6°	3°	8°	5°	2°	4.8
NaiveBayes	1°	1°	1°	8°	1°	2.4
KNeighbors	9°	9°	9°	1°	9°	7.4
NeuralNet	5°	7°	4°	4°	6°	5.2
SVectorMachine	4°	5°	5°	7°	7°	5.6
BaggingTree	7°	6°	2°	6°	4°	5.0
BoostedTree	8°	4°	7°	2°	5°	5.2
Stacking	2°	2°	6°	9°	3°	4.4

Tabla 5.6: Posición según f1-score obtenido. Estudio de deserción al 2do año

F1-SCORE: El modelo de Naive Bayes se posiciona con el mejor f1-score en 4 de los 5 conjuntos de atributos, por otra parte, se puede destacar el rendimiento del modelo de Logistic Regresion y Stacking Generalizer. Sin embargo, se posicionan favorablemente ambos solo en 3 conjuntos, mientras que en los restantes sus rendimientos de f1-score son inferiores.

El modelo de KN Neighbors se posiciona último en 4 de los 5 conjuntos.

DESERCION 2DO AÑO	ACCURACY					
MODELO	C1	C2	C3	C4	C5	POSICIÓN PROMEDIO
Logistic Regresion	4°	8°	4°	4°	7°	5.4
DecisionTree	5°	3°	3°	3°	2°	3.2
NaiveBayes	1°	1°	1°	8°	1°	2.4
KNeighbors	9°	9°	9°	1°	9°	7.4
NeuralNet	7°	7°	5°	5°	8°	6.4
SVectorMachine	3°	6°	6°	7°	6°	5.6
BaggingTree	6°	5°	2°	6°	3°	4.4
BoostedTree	8°	4°	8°	2°	4°	5.2
Stacking	2°	2°	7°	9°	5°	5.0

Tabla 5.7: Posición según accurac obtenido. Estudio de deserción al 2do año

ACCURACY: Se reconoce nuevamente al modelo de Naive Bayes con un rendimiento superior en todos los conjuntos, menos en el cuarto. De manera contraria ocurre que el modelo de KN Neighbors tiene rendimientos inferiores, en todos los conjuntos, menos en el cuarto conjunto.

DESERCION 2DO AÑO	PRECISION					
MODELO	C1	C2	C3	C4	C5	POSICIÓN PROMEDIO
Logistic Regresion	4°	8°	3°	3°	8°	5.2
DecisionTree	5°	3°	7°	4°	2°	4.2
NaiveBayes	1°	1°	1°	8°	1°	2.4
KNeighbors	9°	9°	9°	1°	9°	7.4
NeuralNet	6°	7°	4°	5°	6°	5.6
SVectorMachine	3°	5°	5°	7°	7°	5.4
BaggingTree	7°	6°	2°	6°	3°	4.8
BoostedTree	8°	4°	8°	2°	5°	5.4
Stacking	2°	2°	6°	9°	4°	4.6

Tabla 5.8: Posición según precision obtenido. Estudio de deserción al 2do año

PRECISION: Las posiciones alcanzadas por los modelos presentan un alto nivel de varianza, sin embargo se puede reconocer al modelo de Naive Bayes con un buen rendimiento de precision, mientras que para el modelo de KN Neighbors se reconoce su mal rendimiento.

DESERCION 2DO AÑO	RECALL					
MODELO	C1	C2	C3	C4	C5	POSICIÓN PROMEDIO
Logistic Regresion	3°	2°	2°	1°	3°	2.2
DecisionTree	8°	8°	7°	8°	8°	7.8
NaiveBayes	9°	9°	9°	4°	9°	8.0
KNeighbors	1°	3°	1°	9°	6°	4.0
NeuralNet	2°	1°	3°	3°	1°	2.0
SVectorMachine	6°	4°	4°	2°	4°	4.0
BaggingTree	4°	7°	8°	6°	5°	6.0
BoostedTree	5°	6°	6°	7°	7°	6.2
Stacking	7°	5°	5°	5°	2°	4.8

Tabla 5.9: Posición según recall obtenido. Estudio de deserción al 2do año

RECALL: Se destacan los rendimientos en recall obtenidos por los modelos de Logistic Regresion y Neural Net, ambos ubicandose en la parte alta de las posiciones para cada uno de los conjuntos de atributos. Mientras que los modelos de Decision Tree y KN Neighbors se posicionan con malos resultados.

A continuación se muestra las posiciones promedio obtenidas por los modelo de clasificación bajo las diferentes dimensiones de evaluación.

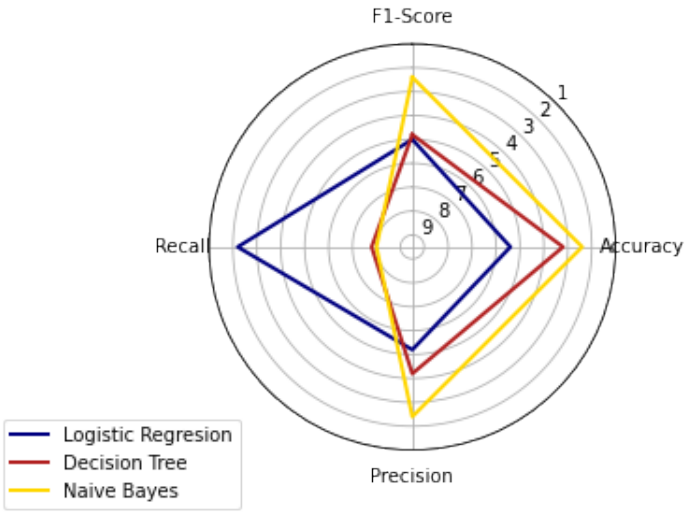
	POSICIÓN PROMEDIO			
MODELO	F1SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	5.0	5.4	5.2	2.2
DecisionTree	4.8	3.2	4.2	7.8
NaiveBayes	2.4	2.4	2.4	8.0
KNeighbors	7.4	7.4	7.4	4.0
NeuralNet	5.2	6.4	5.6	2.0
SVectorMachine	5.6	5.6	5.4	4.0
BaggingTree	5.0	4.4	4.8	6.0
BoostedTree	5.2	5.2	5.4	6.2
Stacking	4.4	5.0	4.6	4.8

Tabla 5.10: Resumen de posición según métrica de desempeño, deserción al 2do año

Se procede a realizar la representación gráfica del cuadro comparativo a través de un gráfico radial.

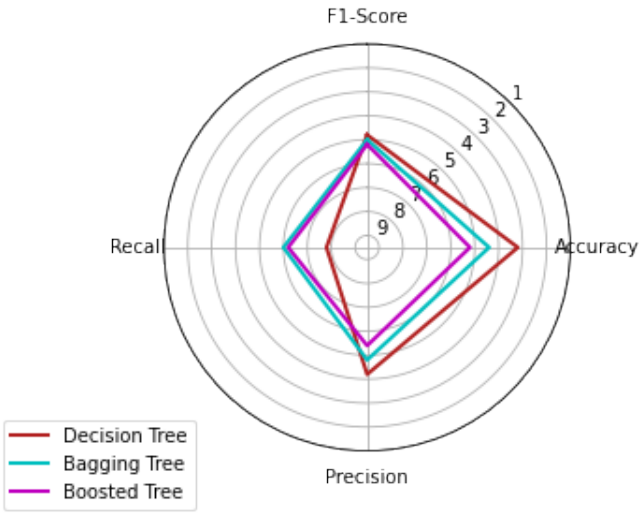
El primer gráfico corresponde solo a los modelos que muestran un rendimiento destacable sobre los demás, el segundo gráfico corresponde a la comparación de los modelos de Decision Tree, Baggging Tree y Boosted Tree. Por último, el tercer gráfico permite comparar al modelo ensamblado de Stacking Generalizer en conjunto a sus modelos utilizados como base.

Figura 5.4: Resultados destacados para el estudio de deserción al segundo año



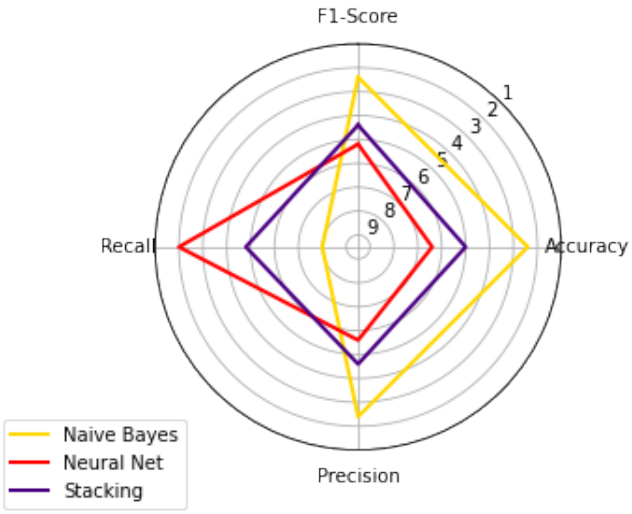
El modelo de Naive Bayes se sitúa como el mejor modelo de la cartera para explicar el fenómeno de deserción al segundo año, mientras que los modelos de Logistic Regresion y Decision Tree tienen un rendimiento medio, esto evaluando sobre la dimensión de f1-score. Se destaca que a pesar de esto, el modelo de Naive Bayes se muestra como mejor opción en cada una de las dimensiones estudiadas al ser comparado con el modelo de Decision Tree. Sin embargo, esto no ocurre con el modelo Logistic Regresion, el cual sobresale por su resultado en recall, permitiendo identificar un mayor numero de estudiantes desertores.

Figura 5.5: Decision Tree y métodos de ensamblaje para el estudio de deserción al segundo año



Se observa que aplicar modelos de ensamblados al método de Decision Tree no genera una mejor posición con respecto a los demás modelos.

Figura 5.6: Comparación de modelo Stacking Generalizer para el estudio de deserción al segundo año



Se tiene que el modelo de Stacking Generalizer, el cual hace uso de los modelos Naive Bayes y Neural Net, se posiciona de manera intermedia entre sus modelos base, lo que permite inferir que logra combinar sus cualidades predictivas.

5.3. Caso de Estudio de Egreso Oportuno

A continuación se presenta la tabla de posiciones sobre los diferentes conjuntos relevantes identificados al estudiar el fenómeno de egreso oportuno. Estos conjuntos se componen de la siguiente forma.

C1: 'EDAD DE INGRESO', 'PREFERENCIA DE POSTULACION', 'PROM. NOTA ESEÑANZA MEDIA', 'PTJE. PSU ELECTIVO'

C2: 'PROM. NOTA ESEÑANZA MEDIA', 'REGION', 'TIPO DE EDUCACION'

C3: 'PTJE. PONDERADO', 'PROM. NOTA ESEÑANZA MEDIA'

C4: 'PTJE. PONDERADO', 'PROM. NOTA ESEÑANZA MEDIA', 'PTJE PSU ELECTIVO', 'REGION', 'TIPO DE EDUCACION'

C5: 'PTJE. N.E.M.', 'PTJE. PSU ELECTIVO', 'PTJE. PSU LENGUAJE', 'REGION', 'TIPO DE EDUCACION'

C6: 'EDAD DE INGRESO', 'PROM. NOTA ENSEÑANZA MEDIA', 'PROVINCIA', 'PTJE. PSU ELECTIVO'

EGRESO OPORTUNO	F1-SCORE						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	2°	6°	5°	7°	7°	1°	4.7
DecisionTree	9°	2°	8°	8°	8°	8°	7.2
NaiveBayes	5°	8°	1°	9°	9°	4°	6.0
KNeighbors	8°	9°	6°	3°	4°	9°	6.5
NeuralNet	1°	5°	3°	4°	1°	3°	2.8
SVectorMachine	3°	7°	4°	5°	3°	5°	4.5
BaggingTree	6°	3°	7°	2°	5°	7°	5.0
BoostedTree	7°	1°	9°	1°	6°	6°	5.0
Stacking	4°	4°	2°	6°	2°	2°	3.3

Tabla 5.11: Posición según f1-score obtenido. Estudio de egreso oportuno

F1-SCORE: Los modelos que más destacan por sobre el resto corresponden a los modelos de Neural Net y Stacking Generalizer, mostrando ambos posiciones correspondiente a la parte media-alta. Por otra parte, se tiene un mal rendimiento del modelo de Decision Tree ubicándose entre la octava y novena posición para 5 de los 6 conjuntos estudiados.

EGRESO OPORTUNO	ACCURACY						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	5°	2°	3°	4°	6°	4°	4.0
DecisionTree	1°	7°	4°	5°	3°	1°	3.5
NaiveBayes	9°	9°	8°	9°	9°	9°	8.8
KNeighbors	2°	1°	1°	3°	2°	3°	2.0
NeuralNet	7°	5°	6°	7°	4°	6°	5.8
SVectorMachine	6°	3°	7°	6°	7°	7°	6.0
BaggingTree	3°	6°	2°	1°	1°	5°	3.0
BoostedTree	4°	8°	5°	2°	5°	2°	4.3
Stacking	8°	4°	9°	8°	8°	8°	7.5

Tabla 5.12: Posición según accuracy obtenido. Estudio de egreso oportuno

ACCURACY: El modelo de KN Neighbors muestra buenos resultados en todos los conjuntos de atributos, destacando claramente sobre el resto, de manera opuesta el modelo Naive Bayes destaca por ubicarse en las peores posiciones.

EGRESO OPORTUNO	PRECISION						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	1°	6°	1°	4°	8°	1°	3.5
DecisionTree	8°	3°	8°	7°	6°	7°	6.5
NaiveBayes	9°	9°	3°	9°	9°	8°	7.8
KNeighbors	6°	8°	5°	3°	2°	9°	5.5
NeuralNet	2°	4°	2°	6°	3°	2°	3.2
SVectorMachine	3°	7°	4°	5°	4°	3°	4.3
BaggingTree	4°	5°	7°	1°	1°	6°	4.0
BoostedTree	7°	2°	9°	2°	7°	5°	5.3
Stacking	5°	1°	6°	8°	5°	4°	4.8

Tabla 5.13: Posición según precision obtenido. Estudio de egreso oportuno

PRECISION: Debido al la gran variabilidad observada en las posiciones encontradas, se dificulta poder concluir la supremacía de un modelo sobre otro, a pesar de esto, los resultados obtenidos por el modelo de Neural Net obtiene resultados medios y altos para cada uno de los conjuntos.

EGRESO OPORTUNO	RECALL						
MODELO	C1	C2	C3	C4	C5	C6	POSICIÓN PROMEDIO
Logistic Regresion	5°	8°	5°	7°	5°	5°	5.8
DecisionTree	9°	3°	8°	9°	8°	9°	7.7
NaiveBayes	1°	1°	2°	1°	1°	1°	1.2
KNeighbors	8°	9°	7°	6°	6°	8°	7.3
NeuralNet	3°	5°	3°	3°	4°	3°	3.5
SVectorMachine	4°	7°	4°	4°	3°	4°	4.3
BaggingTree	7°	4°	6°	8°	9°	7°	6.8
BoostedTree	6°	2°	9°	5°	7°	6°	5.8
Stacking	2°	6°	1°	2°	2°	2°	2.5

Tabla 5.14: Posición según recall obtenido. Estudio de egreso oportuno

RECALL: De destaca los resultados obtenidos por los modelos de Naive Bayes, siendo este superior a todos en las diferentes conjuntos, menos en el tercer conjunto de atributos (puntaje ponderado y promedio de enseñanza media) donde es superado solo por el modelo Stacking Generalizer, en cuanto a este modelo, también se muestra con rendimientos superiores en 5 de los 6 conjuntos de atributos.

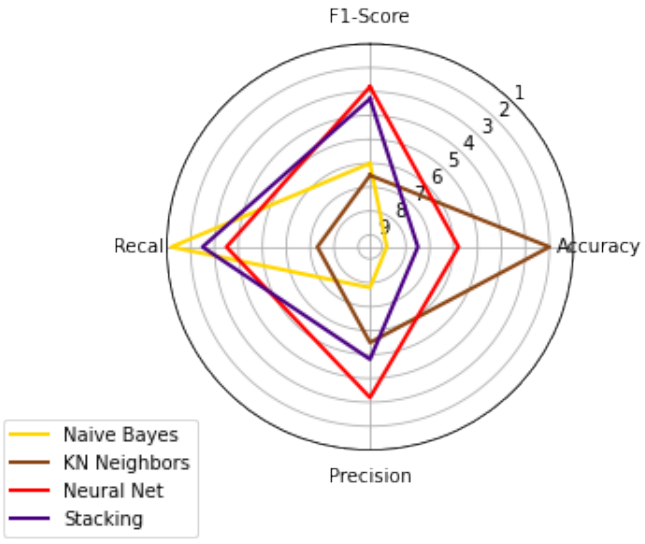
	POSICIÓN PROMEDIO			
MODELO	F1SCORE	ACCURACY	PRECISION	RECALL
Logistic Regresion	4.7	4.0	3.5	5.8
Decision Tree	7.2	3.5	6.5	7.7
Naive Bayes	6.0	8.8	7.8	1.2
KN Neighbors	6.5	2.0	5.5	7.3
Neural Net	2.8	5.8	3.2	3.5
S Vector Machine	4.5	6.0	4.3	4.3
Bagging Tree	5.0	3.0	4.0	6.8
Boosted Tree	5.0	4.3	5.3	5.8
Stacking	3.3	7.5	4.8	2.5

Tabla 5.15: Resumen de posición según métrica de desempeño, egreso oportuno

Si bien es posible obtener resultado de posiciones media, como se vio anteriormente, debido los variadas posiciones que toman los modelos según el conjunto, es difícil asegurar el rendimiento de uno sobre otro.

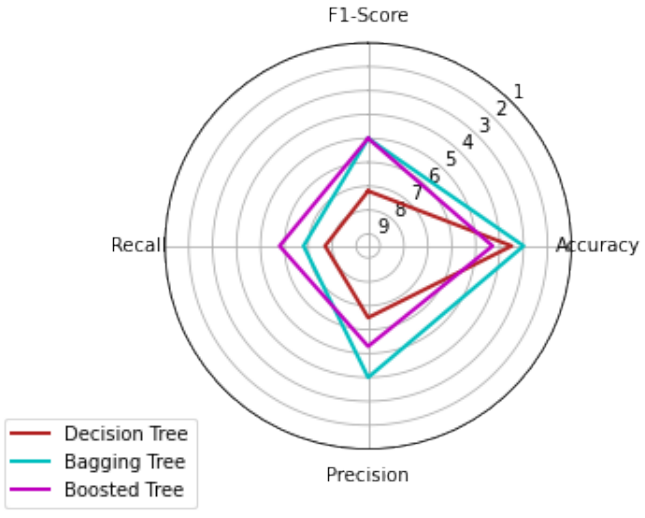
El primer gráfico corresponde solo a los modelos, que a pesar de lo anterior, logran destacar. El segundo gráfico corresponde a la comparación de los modelos de Decision Tree, Baggging Tree y Boosted Tree, correspondiente a métodos ensamblados que utilizan como base al método de Decision Tree. Por último, el tercer gráfico permite comparar al modelo ensamblado de Stacking Generalizer en conjunto a sus modelos utilizados como base.

Figura 5.7: Resultados destacados para el estudio de egreso oportuno



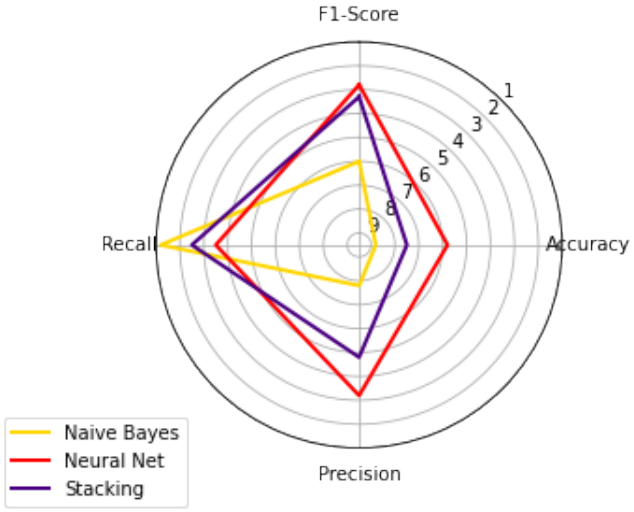
La métrica más importante para reconocer un buen rendimiento del modelo corresponde a f1-score, en tal sentido destacan los modelos de Neural Net y Staking Generalizer, mostrando resultados similares en las distintas dimensiones que se evalúan. Los modelos de Naive Bayes y KN Neighbors obtienen malos rendimientos con respecto a los demás modelos, el primero destaca de manera global con respecto a la métrica recall, sin embargo, no muestra buenos resultados en precision, de manera similar sucede con el modelo KN Neighbors, pero este destaca en metrica accuracy en vez de recall.

Figura 5.8: Decision Tree y métodos de ensamblaje para el estudio de egreso oportuno



Se muestra que al aplicar los métodos de ensamblado al modelo de Decision Tree para explicar el fenómeno de egreso oportuno, se obtienen mejores resultados en cada una de las dimensiones evaluadas.

Figura 5.9: Comparación de modelo Stacking Generalizer para el estudio de egreso oportuno



Se tiene que los resultados obtenidos por el modelo de Stacking Generalizer logra utilizar la información proporcionada por sus modelos bases de nivel 0, Naive Bayes y Neural Net, logrando obtener buenos resultados aceptables en f1-score, precision y recall.

Capítulo 6

Comparación con Estudios Previos

El presente trabajo se desarrolla utilizando como base la metodología propuesta por las investigaciones realizadas en otras universidades, sin embargo, a diferencia de estas, se agregan nuevos puntos y se omiten otros. Por lo tanto, se procede a comparar los resultados en donde ambas investigaciones comparten el mismo procedimiento, con el fin de exponer resultados comparables. Sin embargo, se presenta la salvedad para el caso de egreso oportuno, donde las otras universidades consideran un semestre de holgura, mientras que para la clasificación realizada en el trabajo se hace imposible por restricción de los datos.

A continuación se presentan resultados obtenidos de la investigación realizada en las universidades Adolfo Ibáñez, de Talca, Autónoma y de Chile.

6.1. Selección de Atributos

Con los datos de cada universidad se procede hacer una selección de atributos de igual forma como se realiza en el presente trabajo, donde por medio de la técnica de forward variable selection se obtiene un conjunto propio para cada modelo de clasificación, a continuación se presentan los resultados de las distintas casas de estudio.

6.1.1. Deserción al Primer Año

Universidad Adolfo Ibañez

El atributo más veces seleccionado corresponde al modulo electivo de la PSU, el cual señala si el estudiante rinde la prueba de ciencias y/o historia, seguido por el atributo correspondiente al modulo electivo dentro de la prueba de ciencias. El puntaje que obtiene en la PSU electiva también corresponde a un atributo relevante, afirmando la importancia que tiene la prueba electiva dentro de los estudiantes de la Universidad Adolfo Ibañez. Otro aspecto que también toma relevancia, en menor medida, corresponde a los resultados obtenidos en su etapa previa al ingreso, información capturada por el atributo NEM.

MODELO	ATRIBUTOS SELECCIONADOS
Logistic Regresion	NEM - PTJE PSU ELECTIVO - PREFERENCIA
Decision Tree	PTJE PONDERADO - MODULO PSU CIENCIAS - NEM
Naive Bayes	NEM - PTJE PSU MATEMATICAS - MODULO ELECTIVO
S Vector Machine	PTJE PONDERADO - MODEULO PSU CIENCIAS - MODULO ELECTIVO
Neural Network	MODULO PSU CIENCIAS - PTJE PSU ELECTIVO - MODULO ELECTIVO
KN Neighbors	MODULO ELECTIVO

Tabla 6.1: Selección de atributos, C.E. Deserción 1er año, Universidad Adolfo Ibañez

Universidad de Talca

Para explicar el fenómeno se considera importante la carrera que cursaba el estudiante, siendo este un factor seleccionado por 5 de los 6 modelos, además entre los atributos seleccionados se destacan los correspondientes al puntaje PSU en la prueba de matemáticas y lenguaje. En menor grado, se observa la relevancia de las notas de enseñanza media y de factores socioeconómicos.

MODELO	ATRIBUTOS SELECCIONADOS
Logistic Regresion	CARRERA - PTJE PSU MATEMATICAS PROM. NOTA ENSEÑANZA MEDIA
Decision Tree	CARRERA - PTJE PSU MATEMATICAS - PROVINCIA PTJE PSU LENGUAJE - INGRESO FAMILIAR
Naive Bayes	CARRERA - PTJE PSU MATEMATICAS - DEPENDENCIA EDUCACIONAL
S Vector Machine	CARRERA - PTJE PSU PPS - PTJE PSU PROMEDIO PTJE PSU LENGUAJE - GRUPO INGRESO
Neural Network	CARRERA - PROM. NOTA ENSEÑANZA MEDIA - PTJE PSU MATEMATICAS PTJE PSU LENGUAJE - PTJE PSU PROMEDIO - GRUPO INGRESO
KN Neighbors	NIVEL EDUC PADRES

Tabla 6.2: Selección de atributos, C.E. Deserción 1er año, Universidad de Talca

Universidad Autónoma

Se destaca la importancia del beneficio estatal de gratuidad en los alumnos, siendo el atributo que más se repite junto con el promedio de enseñanza media. También tiene relevancia el puntaje PSU de lenguaje y datos demográficos, como edad de ingreso, genero y ciudad.

MODELO	ATRIBUTOS SELECCIONADOS
Logistic Regresion	PROM. NOTA ENSEÑANZA MEDIA - GRATUIDAD PTJE PSU LENGUAJE - CIUDAD - GENERO - AÑO SABATICO
Decision Tree	GRATUIDAD - PTJE PSU LENGUAJE - PROMEDIO EDAD DE INGRESO
Naive Bayes	PROM. NOTA ENSEÑANZA MEDIA - GRATUIDAD PTJE PSU LENGUAJE - CIUDAD
S Vector Machine	PTJE PSU MATEMATICAS - PROM. NOTA ENSEÑANZA MEDIA AÑO SABATICO - CIUDAD - EDAD DE INGRESO - GRATUIDAD
Neural Network	PROM. NOTA ENSEÑANZA MEDIA
KN Neighbors	PROM NOTA ENSEÑANZA MEDIA - GRATUIDAD - CIUDAD - PTJE PSU LENGUAJE

Tabla 6.3: Selección de atributos, C.E. Deserción 1er año, Universidad Autónoma

Universidad de Chile

Los atributos con mayor presencia dentro de los conjuntos seleccionados corresponden a la edad de ingreso y región, en menor medida se cuenta con los atributos de género y provincia, haciendo notar la importancia de los datos demográficos, al igual que en la Universidad Autónoma. Por otra parte, los estudios de las otra universidades muestran que factores socio-económicos son seleccionados de manera recurrente, sin embargo esto no sucede en el caso de la Universidad de Chile, donde el único dato correspondiente a esta dimensión es el ingreso medio por comuna, el cual no es seleccionado en ninguna ocasión.

MODELO	ATRIBUTOS SELECCONADOS
Logistic Regresion	'ADMISION ESPECIAL' - 'EDAD DE INGRESO' - 'GENERO' - 'REGION'
Decision Tree	'ADMISION ESPECIAL' - 'EDAD DE INGRESO' - 'REGION'
Naive Bayes	'EDAD DE INGRESO' - 'GENERO'
S Vector Machine	'EDAD DE INGRESO' - 'PTJE N.E.M.' - 'REGION'
Neural Net	'EDAD DE INGRESO' - 'PTJE. PSU DE LENGUAJE' - 'REGION'
KN Neighbors	'EDAD DE INGRESO' - 'GENERO' - 'PROVINCIA' - 'REGION'

Tabla 6.4: Selección de atributos, C.E. Deserción 1er año, Universidad de Chile

6.1.2. Deserción al Segundo Año

De igual forma, a través de la técnica 'forward variable selection', para cada universidad se hace una selección de atributos utilizando distintos métodos de clasificación para el estudio de

deserción al segundo año. A continuación se muestran los resultados las otras universidades, siendo estos comparados con los propios de la investigación.

Universidad Adolfo Ibañez

Como es de esperar, se tiene gran importancia el atributo correspondiente al numero de ramos reprobados en el primer año académico, sumado a esto se aprecia la aparición del atributo del promedio de notas en las materias cursadas durante este mismo periodo.

MODELO	ATRIBUTOS SELECCIONADOS
Logistic Regresion	RAMOS REPROB 1ER AÑO - PTJE PSU ELECTIVO
Decision Tree	RAMOS REPROB 1ER AÑO - NIVEL EDUC PADRE - PROM. NOTA ENSEÑANZA MEDIA
Naive Bayes	RAMOS REPROB 1ER AÑO - PTJE RANKING - REGION - RANGO INGRESO
KN Neighbors	PTJE PSU MATEMATICAS - DEPENDENCIA EDUCACIONAL
Neural Network	RAMOS REPROB 1ER AÑO - NIVEL EDUC MADRE - PROMEDIO 1ER AÑO

Tabla 6.5: Selección de atributos, C.E. Deserción 2do año, Universidad Adolfo Ibañez

Universidad de Talca

Nuevamente el número de ramos reprobados en el primer año aparece constantemente entre los conjuntos de atributos.

MODELO	ATRIBUTOS SELECCIONADOS
Logistic Regresion	RAMOS REPROB 1ER AÑO - PTJE PSU ELECTIVO
Decision Tree	RAMOS REPROB 1ER AÑO
Naive Bayes	RAMOS REPROB 1ER AÑO - CARRERA
KN Neighbors	PTJE PSU ELECTIVO
Neural Network	RAMOS REPROB 1ER AÑO - CARRERA - PTJE NEM

Tabla 6.6: Selección de atributos, C.E. Deserción 2do año, Universidad de Talca

Universidad de Chile

Al igual que para la propia investigación, el atributo correspondiente al número de ramos reprobados durante el primer año se presenta como un atributo importante para explicar el

fenómeno, donde el resultado de todas las universidades coinciden en este aspecto. Se destaca también la presencia de atributos correspondientes a las notas obtenidas durante la enseñanza media en cada estudio.

MODELO	ATRIBUTOS SELECCONADOS
Logistic Regresion	'EDAD DE INGRESO' - 'TIPO DE EDUCACION'
Decision Tree	'EDAD DE INGRESO' - 'RAMOS REPROB 1ER AÑO'
Naive Bayes	'EDAD DE INGRESO'
Neural Net	'ADMISION ESPECIAL' - 'EDAD DE INGRESO', 'RAMOS REPROB 1ER AÑO'
KN Neighbors	'COMUNA' - 'RAMOS REPROB. 1ER AÑO'

Tabla 6.7: Selección de atributos, C.E. Deserción 2do año, Universidad de Chile

6.1.3. Egreso Oportuno

Universidad Adolfo Ibañez

El atributo del puntaje PSU ranking destaca sobre el resto de los atributos seleccionados, también se destaca la presencia del promedio de notas de enseñanza media, manifestando la importancia del rendimiento obtenido en la etapa previa educacional.

MODELO	ATRIBUTOS SELECCIONADOS
Logistic Regresion	PTJE RANKING
Decision Tree	PTJE RANKING - PTJE PONDERADO - PTJE PSU LENGUAJE
Naive Bayes	PTJE RANKING - DEPENDENCIA EDUCACIONAL
KN Neighbors	PTJE RANKING - MODULO ELECTIVO
Neural Network	PTJE PSU LENGUAJE - PROM. NOTA ENSEÑANZA MEDIA

Tabla 6.8: Selección de atributos, C.E. Egreso Oportuno, Universidad Adolfo Ibañez

Universidad de Talca

No existe una clara predominancia de algún atributo en particular, así los atributos seleccionados incluyen la carrera cursada; los puntajes PSU de lenguaje y matemáticas; el rendimiento en enseñanza media y características del establecimiento educacional previo.

MODELO	ATRIBUTOS SELECCIONADOS
Logistic Regresion	CARRERA - PTJE PPS - PTJE PSU MATEMATICAS PTJE PSU LENGUAJE - TIPO DE EDUCACION
Decision Tree	PROM. NOTA ENSEÑANZA MEDIA
Naive Bayes	CARRERA - PTJE PPS
KN Neighbors	DEPENDENCIA EDUCACIONAL
Neural Network	PTJE N.E.M.

Tabla 6.9: Selección de atributos, C.E. Egreso Oportuno, Universidad de Talca

Universidad de Chile

Se encuentra similitud con los resultados propios de la investigación y los realizados en otras universidades, donde se concuerda la relevancia del rendimiento obtenido por el estudiante en su etapa educacional previa.

Con respecto a las pruebas de selección, en cada universidad se selecciona al puntaje PSU de lenguaje dentro de los atributos relevantes. Aunque en la Universidad de Chile este es menos recurrente.

MODELO	ATRIBUTOS SELECCONADOS
Logistic Regresion	'EDAD DE INGRESO' - 'PREFERENCIA DE POSTULACION' - 'PROM. NOTA ESEÑANZA MEDIA' - 'PTJE. PSU ELECTIVO'
Decision Tree	'PROM. NOTA ESEÑANZA MEDIA' - 'REGION' - 'TIPO DE EDUCACION'
Naive Bayes	'PTJE. PONDERADO' - 'PROM. NOTA ESEÑANZA MEDIA'
Neural Net	'PTJE. N.E.M.' - 'PTJE. PSU ELECTIVO' - 'PTJE. PSU LENGUAJE' - 'REGION' - 'TIPO DE EDUCACION'
KN Neighbors	'PTJE. PONDERADO' - 'PROM. NOTA ESEÑANZA MEDIA' - 'PTJE PSU ELECTIVO' - 'REGION' - 'TIPO DE EDUCACION'

Tabla 6.10: Selección de atributos, C.E. Egreso Oportuno, Universidad de Chile

6.2. Calidad de Resultados

Dado que los resultados son obtenidos mediante el procedimiento 'forward variable selection', el cual se encuentra limitado en el número de combinaciones de variables evaluadas, se procede a comprobar la calidad de los resultados haciendo un análisis cruzado entre todos los conjuntos de atributos relevantes y el desempeño en f1-score de los modelos.

Si bien el presente proyecto no realiza esta revisión de calidad, si se obtienen estos valores

y por lo tanto se presentan a continuación para comparar los resultados. Estos se muestran a través de una matriz, donde en las filas se encuentra el modelo de clasificación utilizado, mientras que en las columnas se tiene el conjunto de atributos relevantes seleccionado por dicho modelo. De esta forma la diagonal de la matriz representa el rendimiento del modelo al utilizar el conjunto de atributos seleccionado por el mismo. Además se destaca el mejor resultado para cada modelo de clasificación.

6.2.1. Deserción al Primer Año

Universidad Adolfo Ibañez

Se tiene que para los modelos Logistic Regresion, Neural Net y KN Neighbors existe un conjunto de atributos alternativos que proporcionan mejores resultados, sin embargo se tiene que estos valores se encuentran dentro de una desviación estándar del original.

MODELO	Logistic Regresion	Decision Tree	Naive Bayes	SVector Machine	Neural Net	KN Neighbors
Logistic Regresion	.300 ± .056	.261 ± .034	.301 ± .055	.249 ± .031	.227 ± .036	.277 ± .035
Decision Tree	.254 ± .040	.287 ± .043	.258 ± .046	.260 ± .039	.227 ± .036	.262 ± .045
Naive Bayes	.295 ± .051	.264 ± .046	.302 ± .053	.244 ± .033	.227 ± .036	.287 ± .038
SVector Machine	.288 ± .051	.286 ± .052	.285 ± .052	.290 ± .061	.227 ± .036	.274 ± .040
Neural Net	.251 ± .044	.246 ± .045	.247 ± .049	.244 ± .061	.252 ± .046	.284 ± .045
KN Neighbors	.291 ± .057	.284 ± .049	.290 ± .039	.287 ± .037	.275 ± .048	.274 ± .037

Tabla 6.11: F1-Score mediante análisis cruzado, C.E. Deserción al 1er año, U. Adolfo Ibañez

Universidad de Talca

Los modelos obtienen su mejor rendimiento cuando son entrenados con los datos seleccionados al utilizar la técnica de 'forward variable selection'

MODELO	Logistic Regresion	Decision Tree	Naive Bayes	SVector Machine	Neural Net	KN Neighbors
Logistic Regresion	.316 ± .063	.306 ± .063	.307 ± .061	.308 ± .065	.111 ± .017	.308 ± .055
Decision Tree	.243 ± .049	.251 ± .053	.239 ± .065	.202 ± .041	.109 ± .017	.235 ± .046
Naive Bayes	.314 ± .064	.301 ± .056	.317 ± .066	.283 ± .054	.111 ± .016	.295 ± .063
SVector Machine	.275 ± 0.0	.285 ± .058	.241 ± .051	.302 ± .043	.108 ± .018	.283 ± .056
Neural Net	.187 ± 0.0	.176 ± .061	.161 ± .044	.188 ± .059	.223 ± .078	.197 ± .079
KN Neighbors	.307 ± .055	.284 ± .053	.260 ± .047	.298 ± .051	.110 ± .017	.317 ± .055

Tabla 6.12: F1-Score mediante análisis cruzado, C.E. Deserción al 1er año, U. de Talca

Universidad Autónoma

Solo uno de los modelos presenta mejores rendimientos cuando es entrenado con un conjunto de datos alternativos, siendo este el KN Neighbors.

MODELO	Logistic Regresion	Decision Tree	Naive Bayes	SVector Machine	Neural Net	KN Neighbors
Logistic Regresion	.613 ± .187	.424 ± .170	.558 ± .085	.472 ± .098	.558 ± .085	.321 ± .151
Decision Tree	.324 ± .110	.362 ± .146	.324 ± .110	.315 ± .131	.324 ± .110	.244 ± .117
Naive Bayes	.460 ± .098	.353 ± .134	.526 ± .095	.459 ± .081	.526 ± .095	.33 ± .147
SVector Machine	.550 ± .166	.416 ± .121	.465 ± .108	.599 ± .204	.499 ± .138	.286 ± .124
Neural Net	.331 ± .101	.354 ± .146	.500 ± .196	.371 ± .140	.500 ± .196	.351 ± .133
KN Neighbors	.444 ± 0.10	.331 ± .127	.337 ± .109	.359 ± .142	.337 ± .109	.408 ± .187

Tabla 6.13: F1-Score mediante análisis cruzado, C.E. Deserción al 1er año, U. Autónoma

Universidad de de Chile

Se encuentran 3 modelos de clasificación que presentan mejores resultados cuando son entrenados con conjuntos de atributos alternativos, estos corresponden a Logistic Regresion, KN Neighbors y Super Vector Machine, donde en ningún caso el valor obtenido supera en más de una desviación estándar al rendimiento original.

MODELO	Logistic Regresion	Decision Tree	Naive Bayes	SVector Machine	Neural Net	KN Neighbors
Logistic Regresion	.235 ± .032	.228 ± .033	.169 ± .038	.241 ± .038	.244 ± .026	.241 ± .032
Decision Tree	.208 ± .032	.217 ± .032	.185 ± .037	.203 ± .026	.153 ± .023	.210 ± .033
Naive Bayes	.159 ± .034	.192 ± .061	.197 ± .052	.191 ± .041	.184 ± .036	.168 ± .072
SVector Machine	.218 ± .047	.221 ± .052	.178 ± .046	.226 ± .036	.229 ± .034	.236 ± .040
Neural Net	.216 ± .027	.221 ± .033	.168 ± .036	.219 ± .031	.222 ± .033	.216 ± .028
KN Neighbors	.230 ± .040	.205 ± .037	.141 ± .047	.205 ± .018	.157 ± .024	.211 ± .038

Tabla 6.14: F1-Score mediante análisis cruzado, C.E. Deserción al 1er año, U. de Chile

Los niveles de rendimiento difieren significativamente entre los estudios para cada universidad, siendo el presente estudio el que muestran los peores niveles de rendimiento, si bien la diferencia no es tan alta con respecto a las universidades Adolfo Ibañez y de Talca, si lo es al compararse con los resultados en la Universidad Autónoma.

6.2.2. Deserción al Segundo Año

Universidad Adolfo Ibañez

Nuevamente los algoritmos de Logistic Regresion y Neural Net, encuentran un mejor rendimiento al ser entrenados con un conjunto de atributos alternativos, donde este último obtiene un resultado significativamente mejor.

MODELO	Logistic Regresion	Decision Tree	Naive Bayes	Neural Net	KN Neighbors
Logistic Regresion	.468 ± .086	.473 ± .088	.489 ± .084	.175 ± .039	.496 ± .084
Decision Tree	.372 ± .049	.369 ± .069	.381 ± .051	.142 ± 0.03	.388 ± .076
Naive Bayes	.477 ± .091	.488 ± .091	.490 ± .086	.172 ± .040	.474 ± .100
Neural Net	.194 ± .093	.295 ± .134	.271 ± .125	.244 ± .126	.345 ± .093
KN Neighbors	.254 ± .047	.476 ± .075	.466 ± .082	.148 ± .036	.496 ± .065

Tabla 6.15: F1-Score mediante análisis cruzado, C.E. Deserción al 2do año, U. Adolfo Ibañez

Universidad de Talca

3 de los 5 modelos encuentran un mejor rendimiento al ser entrenados con un conjunto de atributos alternativo, Logistic Regresion, Decision Tree y Neural Net. Sin embargo, estos resultados no difieren de manera relevante, ya que se encuentran a menos de una desviación estandar del resultado original.

MODELO	Logistic Regresion	Decision Tree	Naive Bayes	Neural Net	KN Neighbors
Logistic Regresion	.660 ± .021	.652 ± 0.02	.665 ± 0.02	.467 ± .009	.663 ± .028
Decision Tree	.644 ± .020	.644 ± .020	.646 ± .023	.459 ± .007	.646 ± .023
Naive Bayes	.657 ± .028	.652 ± 0.02	.662 ± .027	.467 ± 0.00	.655 ± .021
Neural Net	.598 ± .076	.646 ± .022	.622 ± 0.06	.605 ± .079	.634 ± .062
KN Neighbors	.613 ± .025	.652 ± 0.02	.660 ± .028	.468 ± .006	.664 ± .028

Tabla 6.16: F1-Score mediante análisis cruzado, C.E. Deserción al 2do año, U. de Talca

Universidad de Chile

Los algoritmos de Decision Tree y Neural Net obtienen resultados mejores cuando son entrenados con conjuntos alternativos, donde solo en este último se encuentra una diferencia que supera una desviación estándar del original.

MODELO	Logistic Regresion	Decision Tree	Naive Bayes	Neural Net	KN Neighbors
Logistic Regresion	.221 ± .047	.198 ± .011	.209 ± .037	.199 ± .015	.189 ± .023
Decision Tree	.217 ± .047	.230 ± .026	.206 ± .038	.231 ± .021	.176 ± .024
Naive Bayes	.261 ± .046	.249 ± .031	.276 ± .048	.242 ± .033	.143 ± .030
Neural Net	.217 ± .045	.200 ± .013	.209 ± .037	.202 ± .013	.180 ± .027
KN Neighbors	.126 ± .060	.127 ± .055	.142 ± .071	.162 ± .065	.217 ± .084

Tabla 6.17: F1-Score mediante análisis cruzado, C.E. Deserción al 2do año, U. Chile

El algoritmo de clasificación Neural Net en cada uno de los estudios encuentra un mejor resultado al ser entrenado con un conjunto de atributos alternativo, de tal forma que el algoritmo de selección de atributos no logra obtener una maximización del rendimiento.

Nuevamente se encuentra que los rendimientos obtenidos para el presente estudio corresponden a valores inferiores a los obtenidos en otras universidades.

Capítulo 7

Conclusiones

Los fenómenos de deserción dentro de la FCFM son considerablemente inferiores al resto del sistema educacional expuesto en los informes del SIES (Subsecretaría de Educación Superior), siendo la deserción de primer año tan solo del 5 %, mientras que esta métrica en general oscila en torno al 25 % dentro del sistema de educación superior. También se destaca que el estudiantado se caracteriza por provenir de grupos socioeconómicos altos, presentar buenos resultados al finalizar la etapa educacional previa y hacer ingreso sin interrupciones significativas al sistema educacional chileno.

Con respecto a los algoritmos de clasificación, existen modelos que obtienen rendimientos superiores de manera reiterada cuando son entrenados sobre conjuntos de atributos pertinentes. Por lo tanto, según los interés del investigador es posible realizar las siguientes recomendaciones de modelos de clasificación a instituciones con características similares a la Facultad de Ciencias Físicas y Matemáticas.

- Los resultados obtenidos muestran que los modelos Logistic Regresion, Neural Net y Super Vector Machine destacan sobre el resto al explicar el fenómeno de **deserción al primer año**. Por lo tanto, si el objetivo es identificar de manera precisa quienes van a desertar, ya sea por los costos asociados a la intervención del estudiante, o bien por otras restricciones, se recomienda el uso de Logistic Regresión o Super Vector Machine [1]. Mientras que si se prioriza la identificación anticipada de una cantidad relevante de alumnos que desertarán, se recomienda el uso del modelo Neural Net [2].

  Por otra parte, es posible destacar el rendimiento del modelo Naive Bayes, este muestra resultados de alta precision, si bien no se recomienda su uso debido que deja una gran cantidad de estudiantes sin ser identificados, puede ser de interés al alcanzar una cantidad reducida de estudiantes, pero con alta probabilidad de deserción.
- Los modelos Naive Bayes y Logistic Regresion obtienen buenos resultados para explicar el fenómeno de **deserción al segundo año.** De igual manera, es posible recomendar uno por sobre otro según los intereses del investigador. El modelo Naive Bayes se presenta como un modelo de alta precision, sin dejar de lado el poder predictivo que

[1] Los modelos Logistic Regresion y Super Vector Machine obtienen rendimientos destacables al complementar el análisis con la métrica precision

[2] El modelo Neural Net obtiene un rendimiento destacable al complementar el análisis con la métrica recall

permite alcanzar a una gran cantidad de estudiantes. Por otro lado, el modelo de Logistic Regresion logra identificar a un número mayor de estudiantes que desertarán, sin embargo existe una perdida de precisión considerable.

- Los resultados obtenidos permiten recomendar al modelo de Neural Net para explicar el fenómeno de **egreso oportuno**. Destacando tanto por su precision como su poder predictivo al identificar igualmente a un buen número de estudiantes.

Los modelos de clasificación ensamblados utilizan otros algoritmos de aprendizaje para obtener un mejor rendimiento predictivo utilizando los resultados de algoritmos base. A continuación se presentan las conclusiones de esta afirmación al explicar los fenómenos de deserción al primer año deserción al segundo año y egreso oportuno.

- Para el fenómeno de deserción, tanto al primer año como al segundo año, no es posible asegurar que la afirmación se cumpla. Los métodos Bagging Tree y Boosted Tree, basados en Decision Tree, en general obtiene peores resultados en todos las métricas de rendimiento, menos en recall donde si es posible identificar una mejoría. Sin embargo, al ser aplicados estos métodos en el estudio del egreso oportuno, ambos métodos ensamblados permiten obtener una clara mejoría en las métricas de rendimiento f1-score, recall y precision.
- El estudio también considera la aplicación del método Stacked Generalization, donde nuevamente no es posible asegurar completamente la afirmación. Para el estudio de deserción a primer año los rendimientos se ven fuertemente deteriorados, mientras que en los estudios de deserción al segundo año y egreso oportuno se observa que el rendimiento de este modelo de ensamblado encuentra rendimientos entre los obtenidos por sus modelos base, obteniéndose así un modelo más equilibrado.
- Por lo tanto, si bien es posible alcanzar mejores resultados dado los requerimientos del investigador, esto no siempre se cumple, de modo que se debe planificar mejor manera la estrategia antes de aplicar los métodos de ensamblado, teniendo en cuenta la técnica de ensamblaje y los modelos utilizados como base.

La investigación también considera los resultados obtenidos en otras casas de estudios, al analizar los rendimientos de los modelos de clasificación ejecutados en las distintas universidades se destaca el buen rendimiento del modelo **Logistic Regresion** al explicar el fenómeno de deserción al primer año y del modelo **Naive Bayes** al explicar el fenómeno de deserción al segundo año entre los resultados disponibles, coincidiendo con los resultados encontrados en la presente investigación.

El trabajo también busca identificar los atributos que son relevantes para explicar los fenómenos de estudio, a continuación se presentan las conclusiones halladas a partir de los resultados encontrados en conjunto con los estudios en las otras universidades.

- Los distintos estudios sobre la **deserción al primer año** muestran diferentes tipos de atributos representativos para el fenómeno, si bien las bases de datos no cuentan con las mismas variables, si es posible encontrar semejanzas entre los conjuntos seleccionados. Atributos correspondientes al desempeño en su etapa educacional previa se encuentran dentro de los conjuntos seleccionados en cada una de las casas de estudio.

  Los datos proporcionados por la Universidad de Chile muestran la importancia de

aspectos socio-demográficos para explicar el fenómeno, de igual forma, pero en menor medida, sucede con la Universidad Autónoma, mientras que las universidades Adolfo Ibañez y de Talca se orientan en mayor medida hacia atributos relacionados con los resultados de la PSU.

- Un atributo fundamental para explicar la **deserción al segundo año** corresponde al número de ramos reprobados durante el primer año académico, siendo seleccionado de manera reiterada entre todas las casas de estudio, si bien este hecho no ha de sorprender, gracias al análisis exploratorio realizado en los estudiantes de la Universidad de Chile, es posible presumir la existencia de dos clases distintas de estudiantes, ya que la distribución de este atributo muestra 2 picos, por lo tanto se recomienda profundizar en este asunto con el fin de mejorar los resultados.
- Si bien para los fenómenos anteriores la influencia de los atributos socio-demográficos es notoria para el estudio en la Universidad de Chile, no sucede de igual forma al explicar el **egreso oportuno**, donde se destaca la influencia del promedio de enseñanza media y el puntaje NEM, hecho que se replica en las demás casas de estudio. Dejando en evidencia la relevancia que tienen los resultados obtenidos en la enseñanza media al explicar este fenómeno.
- Se encuentran similitudes entre los atributos obtenidos en las distintas universidades, por otra parte las diferencias no deben de extrañar considerando la diversidad entre los estudiantados presentes en cada una de las casas de estudio.

Los rendimientos obtenidos en el presente trabajo son repetidamente inferiores a los expuestos en las investigaciones de las otras universidades, si bien esto puede deberse a la naturaleza misma de la clase de estudiantes que se inscriben en la facultad, quienes presentan un comportamiento particular, de igual manera se recomienda considerar la incorporación de mayor información a los datos de estudio. Puede ser útil incorporar atributos relacionados con aspectos socio-económicos, como lo son el nivel de ingreso familiar o el nivel educacional de los padres, ya que dentro del análisis exploratorio el ingreso medio por comuna se muestra como un posible buen predictor, también, esta dimensión de atributos es seleccionada dentro de los atributos relevantes en otras universidades y, por último, cuentan con información bibliográfica que respalda esta posición.

Al aplicar la técnica forward variable selection para cada modelo se obtiene un conjunto de atributos relevantes, siendo en general este conjunto el que mejor desempeño proporciona frente a otros conjuntos alternativos seleccionados bajo la misma técnica. Sin embargo, este método sigue probando combinaciones acotadas de atributos, y por lo tanto se recomienda aplicar técnicas más sofisticadas para futuras investigaciones, ya que se observa que la selección de atributos elije una cantidad reducida de atributos, siendo difícil que estos logren incorporar toda la información necesaria para realizar la clasificación correctamente.

Por último, esta técnica utiliza como métrica a maximizar el f1-score, sin embargo esta métrica se considera limitada debido a su construcción, ya que en su formulación utiliza las métricas precision y recall con los mismos pesos. Por lo tanto, según los intereses que tengan las futuras investigaciones puede que esta no sea lo apropiada, pudiendo hacer una combinación de las distintas métricas de rendimiento para priorizar el desempeño de una en particular.

Bibliografía

[1] Servicio de información de educación superior (2019), Informe de retención de primer año de pregrado

[2] Servicio de información de educación superior (2019), Informe duración real y sobreduración de carreras y/o programas, *generación titulados 2014 - 2018*

[3] Mundial, O. B., & OCDE. (2009). La educación superior en Chile. Santiago de Chile.

[4] Donoso, S., & Cancino, V. (2007). Caracterización socioeconómica de los estudiantes de educación superior. Calidad en la Educación, (26), 205-244.

[5] Castillo, J., & Cabezas, G. (2010). Caracterización de jóvenes primera generación en educación superior. Nuevas trayectorias hacia la equidad educativa. Calidad en la Educación, (32), 44-76.

[6] Donoso, S., Donoso, G., & Arias, Ó. (2010). Iniciativas de retención de estudiantes de educación superior. Calidad en la Educación, (33), 15-61.

[7] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of educational research, 45(1), 89-125.

[8] Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. Research in higher education, 12(2), 155-187.

[9] St John, E. P. (2000). The Impact of Student Aid on Recruitment and Retention: What the Research Indicates. New directions for student services, 89, 61-75.

[10] Tierney, W. G. (1999). Models of minority college-going and retention: Cultural integrity versus cultural suicide. Journal of Negro education, 80-91.

[11] González, L. E., Uribe, D., & González, S. O. L. E. D. A. D. (2005). Estudio sobre la repitencia y deserción en la educación superior chilena.

Estudio IESALC, Chile.

[12] Blanco, C., Meneses, F., & Paredes, R. (2018). Más allá de la deserción: trayectorias académicas en la educación superior en Chile. Calidad en la educación, (49), 137-187.

[13] López Guarín, C. E., Gallego Vega, L. E., & Casadiego, M. A. (2016). Implementación de modelos de minería de datos para la definición de tendencias de deserción y permanencia en la universidad nacional de Colombia.

[14] Giovagnoli, P. I. (2002). Determinantes de la deserción y graduación universitaria: una aplicación utilizando modelos de duración. Documentos de trabajo.

[15] Tinto, V. (1989). Definir la deserción: una cuestión de perspectiva. Revista de educación superior, 71(18), 1-9.

[16] Hackman, J. R., & Dysinger, W. S. (1970). Commitment to college as a factor in student attrition. Sociology of Education, 311-324.

[17] Bank, B., Slaving, R., & Biddle, B. (1990). Effects of Peer. Faculty and parental Influences on.

[18] Reay, D., Davies, J., David, M., & Ball, S. J. (2001). Choices of degree or degrees of choice? Class,'race'and the higher education choice process. Sociology, 35(4), 855-874.

[19] Bound, J., Lovenheim, M. F., & Turner, S. (2012). Increasing time to baccalaureate degree in the United States. Education Finance and Policy, 7(4), 375-424.

[20] Ishitani, T. T. (2006). Studying attrition and degree completion behavior among first-generation college students in the United States. The Journal of Higher Education, 77(5), 861-885.

[21] Bell, A., & Valliani, N. (2014). The real cost of college: Time & credits to degree at California State University.

[22] DesJardins, S. L., McCall, B. P., Ahlburg, D. A., & Moye, M. J. (2002). Adding a timing light to the "tool box". Research in Higher Education, 43(1), 83-114.

[23] Adelman, C. (2006). The toolbox revisited: Paths to degree completion from high school through college. US Department of Education.

[24] Fumera, G., & Roli, F. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. IEEE transactions on pattern analysis and machine intelligence, 27(6), 942-956.

[25] Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. Interchange, 1(1), 64-85.

[26] Reyes Rocabado, J., Escobar Flores, C., Duarte Vargas, J., & Ramirez Peradotto, P. (2007). Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. Estudios pedagógicos (Valdivia), 33(2), 101-120.

[27] Stinebrickner, R., & Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. Journal of Labor Economics, 32(3), 601-644.

[28] Jiménez, C. A., Jones, E. A., & Vidal, C. L. (2019). Estudio Exploratorio de Factores que Influyen en la Decisión de la Mujer para Estudiar Ingeniería en Chile. Información tecnológica, 30(4), 209-216.

[29] Rodríguez, A., & Winchester, L. (2001). Santiago de Chile: Metropolización, globalización, desigualdad. EURE (Santiago), 27(80), 121-139.

[30] Gattini, C., Chávez, C., & Albers, D. (2014). Comunas de Chile, según nivel socio-económico, de salud y desarrollo humano. Revisión 2013. Santiago: Observatorio Chileno de Salud Pública y Facultad de Medicina, Universidad de Chile.

[31] Liu, X. Y., Li, Q. Q., & Zhou, Z. H. (2013, December). Learning imbalanced multi-class data with optimal dichotomy weights. In 2013 IEEE 13th International Conference on Data Mining (pp. 478-487). IEEE.

[32] Al-Sudani, S., & Palaniappan, R. (2019). Predicting students' final degree classification using an extended profile. Education and Information Technologies, 24(4), 2357-2369.

[33] Yue, H., & Fu, X. (2017). Rethinking graduation and time to degree: A fresh perspective. Research in Higher Education, 58(2), 184-213.

[34] Swail, W. S. (2003). Retaining Minority Students in Higher Education: A Framework for Success. ASHE-ERIC Higher Education Report. Jossey-Bass Higher and Adult Education Series. Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741.

[35] Lassibille, G., & Gómez, M. L. N. (2011). How long does it take to earn a higher education degree in Spain?. Research in Higher Education, 52(1), 63-80.

[36] Knight, W. E., & Arnold, W. (2000). Towards a Comprehensive Predictive Model of Time to Bachelor's Degree Attainment. AIR 2000 Annual Forum Paper.