

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

**“DISEÑO DE UN PROCESO DE ALERTAS TEMPRANAS PARA DISMINUIR LAS  
DESERCIONES DE LOS ESTUDIANTES DE PRIMER AÑO EN UNA  
INSTITUCIÓN DE EDUCACIÓN SUPERIOR”**

**PROYECTO DE GRADO PARA OPTAR AL GRADO DE MAGÍSTER EN  
INGENIERÍA DE NEGOCIOS CON TECNOLOGÍAS DE INFORMACIÓN**

**FRANCISCA PATRICIA MIRANDA HIDALGO**

**PROFESOR GUÍA:  
EZEQUIEL MUÑOZ KRSULOVIC**

**MIEMBROS DE LA COMISIÓN:  
SEBASTIÁN RIOS PEREZ  
LUCIANO VILLAROEL PARRA**

**SANTIAGO DE CHILE  
2019**



## RESUMEN EJECUTIVO

La deserción estudiantil es un fenómeno que ha ido en aumento durante los últimos años a nivel nacional, por lo tanto, resulta de interés comprender por qué se genera esta situación, que desde la perspectiva de las ciencias sociales corresponde a identificar y comprender cuales son los factores y predictores de la deserción.

El objetivo de este proyecto es identificar cuales son los factores que lleva a los estudiantes a desertar, con el propósito de generar un proceso de alertas tempranas por medio de la creación y aplicación de un modelo que permita detectar tempranamente a los posibles desertores de los estudiantes de pregrado de una Universidad del territorio nacional.

La metodología general del proyecto de rediseño es la Ingeniería de Negocios del doctor Oscar Barros, mientras que para la creación de los modelos predictivos se utiliza la metodología CRISP-DM. El propósito es obtener un modelo que permita ir monitoreando las calificaciones obtenidas por los alumnos en las tres primeras evaluaciones de cada ramo del primer semestre del primer año y así dirigir los esfuerzos de manera localizada, generando un acompañamiento especial en las asignaturas que presentan dificultad.

Los algoritmos utilizados para la creación de los modelos predictivos fueron los siguientes: Árboles de decisión (Gradient Boosted Trees y Random Forest), Regresión Logística, Redes Neuronales y Máquinas de Soporte Vectorial, donde el que presentó los mejores resultados de predicción fue el Random Forest, con un *Accuracy* del 83% y un *Fscore* del 70,3%. Con tal modelo se realizó una prueba piloto a los alumnos que ingresaron a la carrera el año 2017, obteniendo una asertividad del 58,34%.

Se presenta una arquitectura basada en la conexión a la base de datos institucional para extraer los datos de cada asignatura, luego se ejecutan los pasos de la metodología CRISP-DM, finalmente se identifican los casos de uso, diagrama de arquitectura y diagrama de despliegue.

Por último, se realiza una evaluación económica del proyecto, bajo tres escenarios: optimista, conservador y pesimista, utilizando una tasa de descuento del 15%, que es la que utiliza la Universidad para evaluar proyectos para un horizonte de tiempo de 3 años. En todos los escenarios la rentabilidad del proyecto es alta, ya que la inversión inicial es bastante baja, respecto al retorno, que finalmente es la retención de un grupo de alumnos, que en términos económicos significa el pago del arancel.

## DEDICATORIA

*A mis hijas, Fernanda y Francisca  
Gracias por ser mis hijas, mis maestras  
y darme la fuerza para tomar las decisiones correctas  
Confío en que Dios me dará la sabiduría e inteligencia  
para criarlas fuertes, valientes  
y por guiarlas por el buen camino siempre*

## **AGRADECIMIENTOS**

Agradezco a mi familia por el apoyo incondicional que me entregaron durante todo este proceso, gracias por animarme a seguir y darme las fuerzas cuando sentía que me faltaban.

Agradezco a mi madre Ladys Hidalgo Godoy, por su apoyo y contención incondicional. Gracias por tu amor y compañía.

Agradezco de forma especial todo el apoyo a todo el equipo del MBE: a Anita María Valenzuela y Laura Sáez por su ayuda y buena disposición, siempre. A Sebastián Ríos Pérez y Luciano Villarroel Parra por la oportunidad y comprensión. Siempre estaré agradecida de ustedes.

Mis agradecimientos también para Constanza Contreras Piña, por su tiempo y excelente disposición para resolver mis dudas y ayudarme todas las veces que lo necesité

Finalmente, a mi querido profesor guía Ezequiel Muñoz Krsulovic por su dirección durante todo este proceso, excelente disposición para reunirse conmigo las veces que lo necesité y por su preocupación constante por el avance de mi trabajo durante todo este proceso.

También agradezco a la Universidad que me facilitó los datos para trabajar con ellos, que, por motivos de confidencialidad de la información, no puedo mencionar.

# TABLA DE CONTENIDO

<b>CAPÍTULO 1: INTRODUCCIÓN Y CONTEXTO.....</b>	<b>1</b>
1.1 DESCRIPCIÓN GENERAL DE LA INSTITUCIÓN .....	1
1.1.1 <i>Visión</i> .....	1
1.1.2 <i>Misión</i> .....	1
1.1.3 <i>Organigrama</i> .....	1
1.2 OPORTUNIDAD IDENTIFICADA.....	2
1.3 OBJETIVOS Y RESULTADOS ESPERADOS DEL PROYECTO.....	2
1.3.1 <i>Objetivo general</i> .....	2
1.3.2 <i>Objetivos específicos</i> .....	2
1.4 RESULTADOS ESPERADOS .....	2
1.5 ALCANCE .....	3
1.6 RIESGOS POTENCIALES.....	3
<b>CAPÍTULO 2: MARCO TEÓRICO .....</b>	<b>4</b>
2.1 METODOLOGÍA DE INGENIERÍA DE NEGOCIOS.....	4
.....	5
2.2 MODELOS TEÓRICOS DE LA DESERCIÓN .....	5
2.2.1 1970: <i>Spady y su modelo basado en la teoría del suicidio</i> .....	6
2.2.2 1975: <i>Tinto y su modelo basado en la teoría del intercambio</i> .....	7
2.2.3 1985: <i>Bean y su modelo basado en la productividad del ambiente laboral</i> .....	8
2.3 MINERÍA DE DATOS Y LA DESERCIÓN.....	11
2.3.1 <i>Metodología CRISP-DM</i> .....	11
2.3.2 <i>Minería de Datos</i> .....	13
2.3.2.1 Máquinas de aprendizaje .....	15
2.3.2.1.1 Support Vector Machine (SVM).....	15
2.3.2.1.2 Decision Trees (DT) .....	16
2.3.2.1.3 Artificial Neural Network (ANN).....	19
2.3.2.1.4 Logistic Regression (LR).....	22
2.3.2.2 Interpretación y evaluación.....	22
2.3.2.3 Clasificadores .....	24
<b>CAPÍTULO 3: PLANTEAMIENTO ESTRATÉGICO Y MODELO DE NEGOCIOS .....</b>	<b>26</b>
3.1 POSICIONAMIENTO ESTRATÉGICO .....	26
3.2 MAPA ESTRATÉGICO .....	26
3.3 MODELO DE NEGOCIOS .....	26
3.3.1 <i>Propuesta de valor para el cliente</i> .....	27
3.3.2 <i>Recursos claves</i> .....	27
3.3.3 <i>Procesos claves</i> .....	27
3.3.4 <i>Fórmula de utilidades</i> .....	28
3.3.5 <i>Amenazas de nuevos competidores</i> .....	28
3.3.6 <i>Amenazas de productos y servicios sustitutos</i> .....	28
3.3.7 <i>Poder de Negociación de clientes</i> .....	28
3.3.8 <i>Poder de Negociación de Proveedores</i> .....	28
3.3.9 <i>Rivalidad entre Competidores</i> .....	28
3.4 ANÁLISIS FODA .....	29
3.4.1 <i>Fortalezas</i> .....	29
3.4.2 <i>Debilidades</i> .....	29
3.4.3 <i>Oportunidades</i> .....	29
3.4.4 <i>Amenazas</i> .....	29
<b>CAPÍTULO 4: ANÁLISIS DE LA SITUACIÓN ACTUAL.....</b>	<b>30</b>

4.1 ARQUITECTURA DE PROCESOS .....	30
4.1.1 Macroproceso 1: Cadena de Valor.....	30
4.1.2 Macroproceso 2: Desarrollo de nuevas capacidades.....	30
4.1.3 Macroproceso 3: Planificación del Negocio.....	30
4.1.4 Macroproceso 4: Procesos de Apoyo.....	31
4.1.5 Servicios Académicos de Pregrado .....	31
4.1.5.1 Administración de relación con el estudiante.....	31
4.1.5.2 Administración de relación con el académico.....	31
4.1.5.3 Gestión de los Servicios Académicos.....	31
4.1.5.4 Ejecución de los Servicios Académicos .....	32
4.1.5.5 Mantenión de Estado.....	32
4.1.6 Gestión de los Servicios Académicos .....	32
4.1.6.1 Implementación de nuevos programas .....	32
4.1.6.2 Planificación y control de servicios académicos .....	32
4.1.6.3 Decidir entrega de procesos académicos .....	32
4.2 CUANTIFICACIÓN DEL PROBLEMA U OPORTUNIDAD .....	33
4.2.1 Antecedentes: Deserción en la Educación Superior.....	33
4.2.2 Retención de 1er año.....	34
4.2.3 Deserción en la Universidad ACME .....	37
<b>CAPÍTULO 5: PROPUESTA DE DISEÑO DE PROCESOS.....</b>	<b>42</b>
5.1 DIRECCIONES DE CAMBIO Y ALCANCE.....	42
5.2 DISEÑO DETALLADO DE PROCESOS .....	44
5.2.1 Diseño en IDEF0 .....	44
5.3 LÓGICA DE NEGOCIOS .....	45
5.4 PRUEBA DE LA LÓGICA DE NEGOCIOS .....	47
5.4.1 Entendimiento del negocio.....	47
5.4.1.1 Determinar los objetivos del negocio.....	48
5.4.1.2 Evaluación de la situación actual .....	48
5.4.1.3 Determinación de los objetivos de la Minería de Datos.....	48
5.4.2 Entendimiento de los datos.....	48
5.4.2.1 Recopilación de los datos iniciales.....	48
5.4.2.2 Descripción de los datos.....	49
5.4.2.3 Exploración de los datos.....	49
5.4.3 Preparación de los datos.....	57
5.4.4 Modelamiento .....	57
5.4.4.1 Selección de las técnicas de modelamiento.....	58
5.4.4.2 Generar Diseño de Prueba.....	58
5.4.4.3 Construcción de Modelos .....	58
5.4.4.3.1 Árboles de decisión.....	59
5.4.4.3.2 Random Forest.....	60
5.4.4.3.3 Gradient Boosted Trees .....	61
5.4.4.3.4 Support Vector Machine.....	63
5.4.4.3.5 Deep Learning .....	65
5.4.4.3.6 Logistic Regression.....	65
5.4.5 Evaluación.....	67
5.4.5.1 Comparación entre modelos .....	67
5.4.6 Despliegue.....	67
<b>CAPÍTULO 6: PROPUESTA DE APOYO TECNOLÓGICO .....</b>	<b>68</b>
6.1 ESPECIFICACIÓN DE LOS REQUERIMIENTOS .....	68
6.1.1 Requerimientos Funcionales .....	68
6.1.2 Requerimientos No Funcionales .....	68
6.2 ARQUITECTURA TECNOLÓGICA .....	69
6.2 DISEÑO DEL SISTEMA VIRTUAL .....	69
6.2.1 Casos de Uso.....	69

6.2.2	<i>Diagrama de arquitectura del sistema</i> .....	71
6.2.3	<i>Diagrama de despliegue</i> .....	72
<b>CAPÍTULO 7: GESTIÓN DEL CAMBIO</b> .....		<b>73</b>
7.1	CONTEXTO DE LA ORGANIZACIÓN .....	73
7.2	ANÁLISIS DE LOS PRINCIPIOS DE DISEÑO .....	73
7.2.1	<i>Liderazgo y gestión del proyecto de cambio</i> .....	73
7.2.2	<i>Estrategia y sentido del proceso de cambio</i> .....	73
7.2.3	<i>Cambio y Conservación</i> .....	73
7.2.4	<i>Organización y Estructura del proyecto de cambio</i> .....	74
7.2.5	<i>Gestión Emocional</i> .....	74
7.2.6	<i>Comunicaciones</i> .....	74
7.2.7	<i>Desarrollo de Habilidades</i> .....	74
7.2.8	<i>Gestión del Poder</i> .....	74
7.2.9	<i>Monitoreo y Evaluación del Proceso</i> .....	74
7.2.9.10	<i>Inicio, hitos, ritos y cierre</i> .....	75
7.3	CARACTERIZACIÓN DEL CAMBIO .....	75
7.4	FACTORES CRÍTICOS DE ÉXITO .....	75
7.5	PLAN DE GESTIÓN DEL CAMBIO .....	75
<b>CAPÍTULO 8: EVALUACIÓN DEL PROYECTO</b> .....		<b>76</b>
8.1	PLAN PILOTO .....	76
8.2	DEFINICIÓN DE BENEFICIOS Y COSTOS .....	79
<b>CAPÍTULO 9: CONCLUSIONES</b> .....		<b>85</b>
<b>CAPÍTULO 10: BIBLIOGRAFÍA</b> .....		<b>87</b>
<b>CAPÍTULO 11: ANEXOS</b> .....		<b>89</b>



## ÍNDICE DE ILUSTRACIONES

ILUSTRACIÓN 1: ORGANIGRAMA. FUENTE: SE RESERVA LA PRIVACIDAD DE LA INSTITUCIÓN. ....	1
ILUSTRACIÓN 2: MODELO ONTOLÓGICO. FUENTE: BARROS, O. 2017 .....	4
ILUSTRACIÓN 3: METODOLOGÍA DE INGENIERÍA DE NEGOCIOS. FUENTE: BARROS, O. 2017.....	5
ILUSTRACIÓN 4: MODELO PLANTEADO POR SPADY. FUENTE: SPADY, 1970A.....	7
ILUSTRACIÓN 5: MODELO PLANTEADO POR TINTO. FUENTE: TINTO & CULLEN, 1975 .....	8
ILUSTRACIÓN 6: RELACIÓN ENTRE VARIABLES PLANTEADAS POR BEAN. FUENTE: BEAN 1980 .....	10
ILUSTRACIÓN 7: FASES DEL MODELO DE REFERENCIA CRISP-DM.....	11
ILUSTRACIÓN 8: APLICACIONES DE LA MINERÍA DE DATOS. FUENTE: HAN, KAMBER & PEI 2012.....	14
ILUSTRACIÓN 9: "REPRESENTACIÓN GRÁFICA DE LA APLICACIÓN DEL ALGORITMO SVM" .....	16
ILUSTRACIÓN 10: REPRESENTACIÓN GRÁFICA DEL RESULTADO OBTENIDO POR UN ALGORITMO BASADO EN LA MÁQUINA DE APRENDIZAJE ÁRBOL DE DECISIONES" .....	17
ILUSTRACIÓN 11: REPRESENTACIÓN GRÁFICA DE LA SEGMENTACIÓN, SEGÚN ALGORITMO DE ÁRBOL DE DECISIÓN. FUENTE (PROVOST & FAWCETT, 2013) .....	17
ILUSTRACIÓN 12: "DESCRIPCIÓN GRÁFICA DE UN PERCEPTRÓN" .....	19
ILUSTRACIÓN 13: "EJEMPLO DE ESTRUCTURA DE UNA RED NEURONAL ARTIFICIAL".....	21
ILUSTRACIÓN 14: EJEMPLO DE MATRIZ DE CONFUSIÓN.....	23
ILUSTRACIÓN 15: EJEMPLO ERROR DE CLASIFICACIÓN.....	24
ILUSTRACIÓN 16: POSICIONAMIENTO ESTRATÉGICO. FUENTE: ELABORACIÓN PROPIA .....	26
ILUSTRACIÓN 17: MODELO DE NEGOCIOS. FUENTE: ELABORACIÓN PROPIA CON BASE EN JOHNSON 2008.....	27
ILUSTRACIÓN 18: MACROPROCESOS UNIVERSIDAD. FUENTE: ELABORACIÓN PROPIA .....	30
ILUSTRACIÓN 19: SERVICIOS ACADÉMICOS DE PREGRADO. FUENTE: ELABORACIÓN PROPIA .....	31
ILUSTRACIÓN 20: GESTIÓN DE SERVICIOS ACADÉMICOS. FUENTE: ELABORACIÓN PROPIA .....	32
ILUSTRACIÓN 21: TASAS DE RETENCIÓN, PERSISTENCIA EN LA MISMA INSTITUCIÓN Y EN EDUCACIÓN SUPERIOR DE 1º AÑO PARA CARRERAS DE PREGRADO POR TIPO DE INSTITUCIÓN, COHORTE 2017. FUENTE: WWW.MIFUTURO.CL.....	34
ILUSTRACIÓN 22: EVOLUCIÓN DE RETENCIÓN DE 1º AÑO POR TIPO DE INSTITUCIÓN. FUENTE: WWW.MIFUTURO.CL.....	35
ILUSTRACIÓN 23: EVOLUCIÓN DE RETENCIÓN DE 1º AÑO POR TIPO DE CARRERA. FUENTE: WWW.MIFUTURO.CL.....	35
ILUSTRACIÓN 24: EVOLUCIÓN DE RETENCIÓN DE 1º AÑO POR TIPO DE INSTITUCIÓN Y CONDICIÓN DE ACREDITACIÓN INSTITUCIONAL. FUENTE: WWW.MIFUTURO.CL.....	36
ILUSTRACIÓN 27: PLANIFICACIÓN Y CONTROL DE SERVICIOS ACADÉMICOS. FUENTE: ELABORACIÓN PROPIA .....	45
ILUSTRACIÓN 28: DESARROLLO MODELO PREDICTIVO .....	45
ILUSTRACIÓN 29: LÓGICA DE NEGOCIOS PROPUESTA. FUENTE: ELABORACIÓN PROPIA.....	46
ILUSTRACIÓN 30: DIAGRAMA DE CASO DE USO .....	69
ILUSTRACIÓN 31: "ARQUITECTURA TECNOLÓGICA". FUENTE: ELABORACIÓN PROPIA .....	71

## ÍNDICE DE TABLAS

TABLA 1: LISTA DE VARIABLES PLANTEADAS POR BEAN. FUENTE: BEAN 1980 .....	8
TABLA 2: TOTAL DE ESTUDIANTES QUE DESERTAN POR AÑO DE INGRESO. FUENTE: ELABORACIÓN PROPIA .....	37
TABLA 3: TOTAL DE ESTUDIANTES QUE DESERTAN POR GÉNERO. FUENTE: ELABORACIÓN PROPIA .....	37
TABLA 4: ESTUDIANTES QUE DESERTAN SEGÚN TIPO DE DEPENDENCIA. FUENTE: ELABORACIÓN PROPIA .....	37
TABLA 5: ESTUDIANTES QUE DESERTAN SEGÚN PROMEDIO DE PUNTAJES. FUENTE: ELABORACIÓN PROPIA .....	38
TABLA 6: ESTUDIANTES QUE DESERTAN SEGÚN PROMEDIO DE PUNTAJES. FUENTE: ELABORACIÓN PROPIA .....	38
TABLA 7: ESTUDIANTES QUE DESERTAN SEGÚN PROCESO AL CUAL POSTULA. FUENTE: ELABORACIÓN PROPIA .....	38
TABLA 8: TOTAL DE ESTUDIANTES QUE DESERTAN POR EDAD. FUENTE: ELABORACIÓN PROPIA .....	39
TABLA 9: ESTUDIANTES QUE DESERTAN SEGÚN PREFERENCIA .....	39
TABLA 10: ESTUDIANTES QUE DESERTAN RESPECTO A TIPO DE INSTITUCIÓN ANTERIOR. FUENTE: ELABORACIÓN PROPIA .....	39
TABLA 11: ESTUDIANTES QUE DESERTAN SEGÚN ESTADO CIVIL. FUENTE: ELABORACIÓN PROPIA .....	40
TABLA 12: ESTUDIANTES QUE DESERTAN RESPECTO A LA POSESIÓN DE TRABAJO. FUENTE: ELABORACIÓN PROPIA .....	40
TABLA 13: ESTUDIANTES QUE DESERTAN SEGÚN INGRESO BRUTO TOTAL DEL GRUPO FAMILIAR. FUENTE: ELABORACIÓN PROPIA .....	40
TABLA 14: ESTUDIANTES QUE DESERTAN SEGÚN NÚMERO DE PERSONAS QUE COMPONEN EL GRUPO FAMILIAR. FUENTE: ELABORACIÓN PROPIA .....	41
TABLA 15: ESTUDIANTES QUE DESERTAN SEGÚN EL N° DE PERSONAS QUE TRABAJAN EN FORMA REMUNERADA EN EL GRUPO FAMILIAR. FUENTE: ELABORACIÓN PROPIA .....	41
TABLA 16: VARIABLE "ESTRUCTURA DE EMPRESA Y MERCADO" .....	42
TABLA 17: VARIABLE "ANTICIPACIÓN" .....	42
TABLA 18: VARIABLE: "COORDINACIÓN" .....	43
TABLA 19: VARIABLE "PRÁCTICAS DE TRABAJO" .....	43
TABLA 20: VARIABLE "INTEGRACIÓN DE PROCESOS CONEXOS" .....	44
TABLA 21: VARIABLE "MANTENCIÓN CONSOLIDADA DE ESTADO" .....	44
TABLA 22: MALLA CARRERA DISEÑO INDUSTRIAL. FUENTE: ELABORACIÓN PROPIA .....	49
TABLA 23: VALORES PERDIDOS POR ASIGNATURA, 1ER SEMESTRE. FUENTE: ELABORACIÓN PROPIA .....	50
TABLA 24: "ANÁLISIS NOTA PARCIAL 1-DISI6011". FUENTE: ELABORACIÓN PROPIA .....	50
TABLA 25: "ANÁLISIS NOTA PARCIAL 2-DISI6011" FUENTE: ELABORACIÓN PROPIA .....	50
TABLA 26: "ANÁLISIS NOTA PARCIAL 3-DISI6011" FUENTE: ELABORACIÓN PROPIA .....	51
TABLA 27: "ANÁLISIS NOTA PARCIAL 1 - DISI6012". FUENTE: ELABORACIÓN PROPIA .....	51
TABLA 28: "ANÁLISIS NOTA PARCIAL 2 - DISI6012". FUENTE: ELABORACIÓN PROPIA .....	52
TABLA 29: "ANÁLISIS NOTA PARCIAL 3 - DISI6012". FUENTE: ELABORACIÓN PROPIA .....	52
TABLA 30: "ANÁLISIS DE NOTA PARCIAL 1 - DISI6013". FUENTE: ELABORACIÓN PROPIA .....	52
TABLA 31: "ANÁLISIS NOTA PARCIAL 2 - DISI6013". FUENTE: ELABORACIÓN PROPIA .....	53
TABLA 32: "ANÁLISIS NOTA PARCIAL 3 - DISI6013". FUENTE: ELABORACIÓN PROPIA .....	53
TABLA 33: "ANÁLISIS NOTA PARCIAL 1- DISI6014". FUENTE: "ELABORACIÓN PROPIA" .....	53
TABLA 34: "ANÁLISIS NOTA PARCIAL 2 - DISI6014". FUENTE: ELABORACIÓN PROPIA .....	54
TABLA 35: "ANÁLISIS NOTA PARCIAL 3 - DISI6014". FUENTE: ELABORACIÓN PROPIA .....	54
TABLA 36: "ANÁLISIS NOTA PARCIAL 1 - DISI6015". FUENTE: ELABORACIÓN PROPIA .....	54
TABLA 37: "ANÁLISIS NOTA PARCIAL 2 - DISI6015". FUENTE: ELABORACIÓN PROPIA .....	55
TABLA 38: "ANÁLISIS NOTA PARCIAL 3- DISI6015". FUENTE: ELABORACIÓN PROPIA .....	55
TABLA 39: "ANÁLISIS NOTA PARCIAL 1 – HUMI6011". FUENTE: ELABORACIÓN PROPIA .....	55
TABLA 40: "ANÁLISIS NOTA PARCIAL 2 – HUMI6011". FUENTE: ELABORACIÓN PROPIA .....	56
TABLA 41: "ANÁLISIS NOTA PARCIAL 3 - HUMI6011". FUENTE: ELABORACIÓN PROPIA .....	56
TABLA 42: "ANÁLISIS NOTA PARCIAL 1 - PPSB001". FUENTE: ELABORACIÓN PROPIA .....	56
TABLA 43: "ANÁLISIS NOTA PARCIAL 2- PPSB001". FUENTE: ELABORACIÓN PROPIA .....	57
TABLA 44: PERFORMANCE ÁRBOL DE DECISIÓN .....	59
TABLA 45: INDICADORES ÁRBOL DE DECISIÓN .....	60
TABLA 46: PERFORMANCE RANDOM FOREST .....	61
TABLA 47: INDICADORES RANDOM FOREST .....	61
TABLA 48: PERFORMANCE GRADIENT BOOSTED TREES .....	62

TABLA 49: INDICADORES GRADIENT BOOSTED TREES .....	62
TABLA 50: PERFORMANCE SUPPORT VECTOR MACHINE .....	64
TABLA 51: INDICADORES SUPPORT VECTOR MACHINE .....	64
TABLA 52: INDICADORES DEEP LEARNING .....	65
TABLA 53: VALORES REGRESIÓN LOGÍSTICA.....	66
TABLA 54: INDICADORES LOGISTIC REGRESSION .....	66
TABLA 56: PERFORMANCE RANDOM FOREST - PRUEBA PILOTO.....	77
TABLA 57: INDICADORES RANDOM FOREST - PRUEBA PILOTO.....	77
TABLA 58: COMPARACIÓN VALORES RANDOM FOREST .....	77
TABLA 59: SIMULACIÓN NOTAS RANDOM FOREST.....	79
TABLA 60: "PORCENTAJE DE DESERCIÓN DE ALUMNOS EN LA CARRERA DE DISEÑO EN INDUSTRIAL" .....	79
TABLA 61: "CANTIDAD DE ALUMNOS QUE DESERTAN POR AÑO" .....	79
TABLA 62: "CANTIDAD ACUMULADA DE ALUMNOS QUE DESERTAN".....	80
TABLA 63: "ARANCEL DE REFERENCIA POR AÑO" .....	80
TABLA 64: "INGRESOS QUE DEJA DE PERCIBIR LA UNIVERSIDAD POR CONCEPTO DE DESERCIÓN DE ALUMNOS".....	80
TABLA 65: "PROYECCIÓN DE INGRESOS QUE SE DEJAN DE PERCIBIR" .....	82

## ÍNDICE DE GRÁFICOS

GRÁFICO 1: DESEMPEÑO DE LOS PARÁMETROS - ÁRBOLES DE DECISIÓN .....	59
GRÁFICO 2: DESEMPEÑO DE LOS PARÁMETROS - RANDOM FOREST .....	60
GRÁFICO 3: DESEMPEÑO DE LOS PARÁMETROS - GRADIENT BOOSTED TREES.....	62
GRÁFICO 4: DESEMPEÑO DE LOS PARÁMETROS - SVM.....	63

# CAPÍTULO 1: INTRODUCCIÓN Y CONTEXTO

## 1.1 DESCRIPCIÓN GENERAL DE LA INSTITUCIÓN

La Universidad ACME es una institución de educación superior del Estado de Chile, acreditada en 2016 por cuatro años en las áreas de Gestión Institucional, Docencia de Pregrado y Vinculación con el Medio.

La Universidad ofrece un total de 29 carreras de pregrado con ingreso PSU en las áreas del conocimiento de Administración y Economía, Diseño y Arquitectura, Construcción, Ciencias e Ingeniería, en modalidades diurna y vespertina.

Cuenta con tres campus a lo largo del país con más de 62 mil metros cuadrados de construcción entre aulas, laboratorios, bibliotecas, casinos y salones para eventos. Su matrícula total es de 8.432 alumnos y en el proceso de admisión 2016 ingresaron 2.373 nuevos estudiantes, con más del 60% de ellos beneficiados con la gratuidad por pertenecer a los primeros quintiles socioeconómicos de la población

### 1.1.1 Visión

Formar personas con altas capacidades académicas y profesionales, en el ámbito tecnológico, apoyada en la generación, transferencia, aplicación y difusión del conocimiento en las áreas del saber que le son propias, para contribuir al desarrollo sustentable del país y de la sociedad de la que forma parte<sup>1</sup>.

### 1.1.2 Misión

Ser reconocida por la formación de sus egresados, la calidad de su educación continua, la construcción de capacidades y fortalecimiento de la investigación, creación, innovación y transferencia en las áreas del saber, por su cuerpo académico de excelencia y por una gestión institucional que asegura su sustentabilidad e implementación de un sistema integral de calidad en todo su quehacer institucional<sup>2</sup>

### 1.1.3 Organigrama

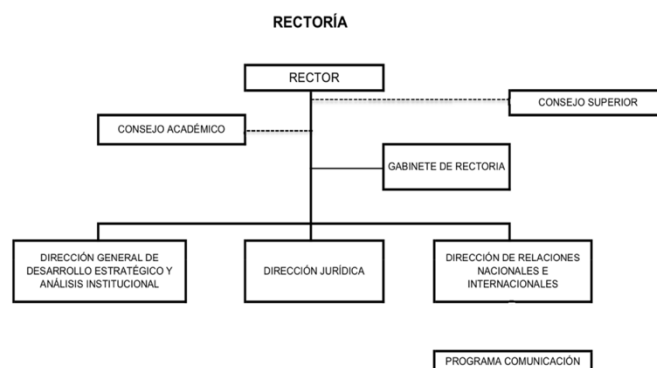


Ilustración 1: Organigrama. Fuente: Se reserva la privacidad de la Institución.

<sup>1</sup> Fuente: PDE 2016-2020

<sup>2</sup> Fuente: PDE 2016-2020

## **1.2 OPORTUNIDAD IDENTIFICADA**

Con el objetivo de orientar constantemente el quehacer universitario a la misión y rol social de esta Institución de Educación Superior, se realiza un levantamiento detallado de la situación actual, identificando las brechas y oportunidades de mejora que existen en determinadas áreas de la Universidad.

Una de las líneas de acción que se determinó como prioritaria tiene relación con el diseño y rediseño de algunos procesos de gestión de docencia, ya que la docencia es uno de los ejes principales dentro de la misión institucional.

Dada la importancia y relevancia que tiene para el aseguramiento de la calidad, así como también, el costo que implica tanto para los alumnos como para la Universidad la deserción estudiantil, es que surge la necesidad de entender la naturaleza del fenómeno y monitorear el proceso a través de la creación de un sistema de Alertas Tempranas.

El diseño, creación e implementación de este proceso, permitirá generar acciones preventivas a lo largo del semestre ya que a través de la identificación de las variables que describen la deserción al interior de la Universidad y de los valores que tomen estas variables, se podrá crear indicadores de gestión que permita crear una alerta real y dirigir los esfuerzos y recursos, sobre un grupo de específico de alumnos.

## **1.3 OBJETIVOS Y RESULTADOS ESPERADOS DEL PROYECTO**

En esta tesis se realizará un estudio del fenómeno de la deserción en el contexto universitario chileno. Los objetivos generales y específicos de esta investigación son:

### **1.3.1 Objetivo general**

Diseñar un proceso que genere alertas e identifique tempranamente a los alumnos de pregrado de primer año con mayor probabilidad de desertar, a través de la construcción de un modelo predictivo que permita detectar a los posibles desertores de pregrado, por medio de la caracterización de una serie de atributos personales, socioeconómicos, grupo familiar y de rendimiento universitario de primer año.

### **1.3.2 Objetivos específicos**

- Identificar los predictores que describen la deserción.
- Aplicar técnicas de minería de datos para generar modelos que permitan predecir la deserción y la no deserción semestral.
- Encontrar el mejor modelo de predicción de deserción, utilizando distintos algoritmos y técnicas.
- Diseñar un proceso de alertas tempranas identificando actores relevantes, reglas de negocio e información relevante, que soporte el modelo de predicción.

## **1.4 RESULTADOS ESPERADOS**

- Implementación de los roles y de las responsabilidades de la nueva unidad de alertas tempranas.

- Generar planes dirigidos de acompañamiento o nivelación a los estudiantes que tienen mayor probabilidad de desertar.
- Aumentar la retención de alumnos en un 10% en las carreras con mayores tasas de deserción.

### 1.5 ALCANCE

El presente proyecto aborda la deserción de los estudiantes de Pregrado de los años 2010 al 2016, de la carrera Diseño Industrial, donde su vía de ingreso ha sido la Prueba de Selección Universitaria (PSU).

El presente estudio considera la deserción que se produce en el primer año, ya que es la más significativa, donde el impacto del primer año para la carrera mencionada anteriormente corresponde a 5,07%, de la matrícula de la cohorte.

### 1.6 RIESGOS POTENCIALES

A continuación, se exponen los 3 riesgos potenciales del proyecto

<b>Tipo</b>	<b>Probabilidad</b>	<b>Impacto</b>	<b>Plan de Mitigación</b>
Inconsistencias en la información	Alta	Aumento de tiempos de procesamiento, mal diagnóstico y resultados	Exigir a la parte técnica que realice correcciones de esto.
Incorrecta selección de variables	Media	Un modelo que no responda a los objetivos	Contrastar resultados con los antiguos modelos.
Gestión del cambio	Media	Resistencia a las nuevas mejoras	Capacitación y acompañamiento del proceso.

## CAPÍTULO 2: MARCO TEÓRICO

### 2.1 METODOLOGÍA DE INGENIERÍA DE NEGOCIOS

En la búsqueda de la competitividad, las empresas se han visto obligadas a buscar ventajas sustentables. De acuerdo con Porter y otros autores, éstas pueden provenir de dos vertientes: la efectividad operacional y la creación de valor único para los clientes. El planteamiento de la Ingeniería de Negocios es que ambas maneras de generar competitividad implican diseñar la estructura y actividades del negocio en forma sistémica y en una combinación única, que las haga difíciles de igualar<sup>2</sup>. Es por esta razón que la Ingeniería de Negocios integra: el Modelo de Negocios, la Estructura Organizacional, los Procesos de Negocios, Sistemas y Tecnologías de Información en la búsqueda del diseño de servicios. Dentro del trabajo de Barros (2017) converge todos estos conceptos y añade una mirada sobre la Arquitectura Empresarial (Enterprise Architecture) en una Ontología para el diseño de Negocios, presentada en la Ilustración 2.

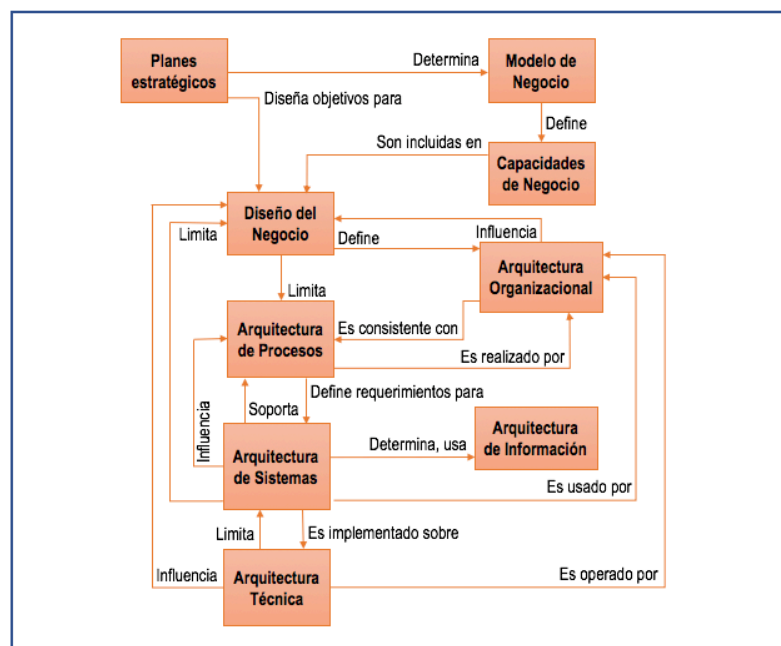


Ilustración 2: Modelo Ontológico. Fuente: Barros, O. 2017

Esta Ontología incorpora la Planificación Estratégica de la empresa con el Modelo de negocios como los principales pilares para el diseño del negocio, donde luego se encuentran las arquitecturas necesarias para el sustento del diseño permitiendo así el funcionamiento de las distintas actividades. Las arquitecturas son las siguientes:

- **Arquitectura de Procesos:** Arquitectura que establece los procesos necesarios para el diseño del negocio, la relación entre estos y la lógica del negocio.
- **Arquitectura Organizacional:** Arquitectura que se relaciona directamente con la de Procesos de acuerdo con la estructura de los roles organizacionales, el trabajo a realizar y por quien.

<sup>2</sup> Barros, O. 2017

- **Arquitectura de Sistemas:** Arquitectura que define el sistema de información que posee la organización y el soporte a los procesos que entrega.
- **Arquitectura de Información:** Arquitectura de la Información que posee la organización en base a sus datos operacionales.
- **Arquitectura Técnica:** Arquitectura relacionada con todo el soporte físico (hardware) de los sistemas de las arquitecturas previamente mencionadas.

Este modelo Ontológico enfatiza el concepto de hacer operativas las ideas y focos establecidos en la estrategia de la organización en base a un alineamiento de ésta con las arquitecturas encargadas del funcionamiento operacional. Es así, como surge el concepto de Metodología de Ingeniería de Negocios que comenta Barros (2016) en el mismo trabajo, la cual detalla las etapas a seguir para el diseño del negocio. La metodología queda detallada en la Ilustración 3.



Ilustración 3: Metodología de Ingeniería de Negocios. Fuente: Barros, O. 2017

## 2.2 MODELOS TEÓRICOS DE LA DESERCIÓN

La deserción universitaria es entendida como la salida, voluntaria o involuntaria, de un alumno del programa de estudio en el que se inscribió. La deserción involuntaria ocurre cuando por decisión institucional el estudiante no puede seguir sus estudios por razones académicas o disciplinarias, mientras que la voluntaria se manifiesta a través de la renuncia formal o del abandono no informado del estudiante (Tinto & Cullen, 1975).

La deserción conlleva costos altísimos a los estudiantes, sobre todo para aquellos que provienen de familias de escasos recursos, los cuales, en general, deben recurrir a créditos bancarios para financiar los estudios superiores. Este costo, no recae solamente sobre los alumnos, sino que sobre todo el sistema educativo, ya que:

- Se genera un congelamiento del financiamiento a la Institución Educativa.



- se pierde una vacante que pudo ser utilizada por otro estudiante que podría haber finalizado el programa.
- se estanca el desarrollo educacional del país, disminuyendo el capital humano avanzado, principalmente en aquellas profesiones mayormente demandadas (Tinto, 2007).

Sin embargo, aun con toda la investigación desarrollada desde distintas disciplinas, el fenómeno de la deserción sigue ocurriendo y son pocas las herramientas que se han generado para mitigar sus efectos negativos. Esto genera una oportunidad para que nuevas disciplinas, principalmente aplicadas, tales como, la ingeniería de negocios y minería de datos, respondan al desafío del mejoramiento de la gestión de la deserción.

Dado lo anterior, surge la necesidad de identificar y analizar los factores que causan este fenómeno al interior de una de las Universidades chilenas, con el propósito de comprender el comportamiento de las variables que describen la deserción y generar medidas preventivas al interior de la organización, permitiendo aumentar la retención estudiantil, y así, mejorar la gestión académica Institucional.

Con el objetivo de entender la evolución de la teoría respecto de la deserción, se mostrarán de manera cronológica tres modelos teóricos que son usados por la mayoría de los investigadores que estudian la deserción:

- Modelo basado en la teoría del suicidio,
- Modelo basado en la teoría del intercambio y
- modelo basado en el modelo de productividad del ambiente laboral.

A continuación, se explicará cada uno de ellos.

### **2.2.1 1970: Spady y su modelo basado en la teoría del suicidio**

Uno de los primeros estudios relacionados con la deserción es el desarrollado por Spady. El autor utiliza los principios del suicidio de Durkheim, el cual establece que la decisión de suicidarse no puede ser explicado solamente por factores individuales, puesto que es un hecho social, ya que es generado por la ruptura del individuo con su sistema social debido a su imposibilidad de integrarse a la sociedad (Durkheim, 1951). Siguiendo esta lógica, Spady establece que la deserción sería un resultado de la no integración del individuo con su entorno educacional y alude a que el entorno familiar y sus características afectan fuertemente al estudiante, ya que estos lo exponen a influencias, expectativas y demandas que podrían afectar tanto en su integración social con sus pares en el ambiente universitario como en el rendimiento académico.

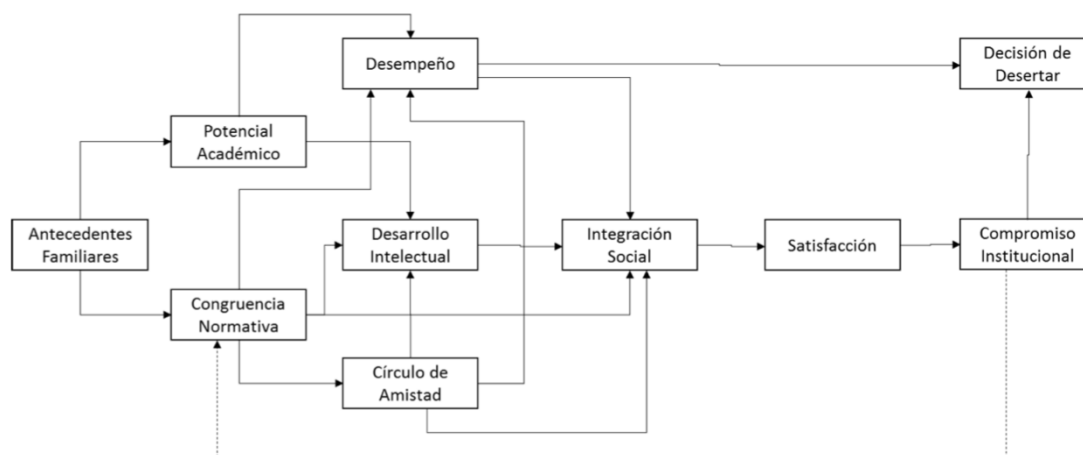


Ilustración 4: Modelo planteado por Spady. Fuente: Spady, 1970a

Tal como se ve en la Ilustración 4, Spady plantea que los antecedentes familiares impactan directamente en el potencial académico y en la congruencia normativa del estudiante, el cual se refiere a la compatibilidad de las actitudes, intereses y disposición personal del individuo con las características del medio. Tanto el potencial académico y la congruencia normativa afectan el desempeño académico, el desarrollo intelectual y su integración con los pares en su ambiente educacional. Cada atributo que defina estos factores impacta directamente a la integración social del individuo, que en consecuencia definirá el nivel de satisfacción y consecuentemente el compromiso con la institución educacional. Todos estos factores tendrán incidencia en la decisión final del estudiante para desertar, ya sea voluntaria o involuntariamente.

### 2.2.2 1975: Tinto y su modelo basado en la teoría del intercambio

En el año 1975 Tinto realiza una revisión de los modelos desarrollados hasta entonces respecto de la deserción. Dentro de esta revisión, resalta el trabajo realizado por Spady y siete años más tarde complementa su modelo incorporando la teoría del intercambio desarrollado por Nye. La teoría del intercambio plantea que los seres humanos evitan aquellas conductas que les generan costos de algún tipo y buscan beneficios en las relaciones, interacciones y estados emocionales que generan con sus pares y la institución educacional (Nye, 1976). Bajo esta perspectiva, para Tinto los estudiantes se mantendrían en el programa que se inscribieron siempre y cuando los beneficios percibidos superen el esfuerzo, dedicación y otros costos personales; y si existe alguna otra actividad que le genere mayores beneficios la decisión final del estudiante podría desencadenar en una deserción (Tinto, 1982).

Según el modelo mostrado en la Ilustración 5, se desprende que cuando un estudiante ingresa a un programa de educación superior, este plantea inicialmente sus compromisos con la institución y con sus objetivos personales de obtener un grado en la institución. Tales compromisos serán afectados por sus antecedentes familiares, como por ejemplo nivel sociocultural; por sus atributos personales, como edad y género; y por su experiencia académica preuniversitaria. Luego de un tiempo razonable estando en el programa, el estudiante reevaluará sus compromisos iniciales de acuerdo con su

integración social y su desempeño académico en la institución, cuyos efectos podrían desencadenar la deserción si el estudiante percibe que los costos sean mayores que los beneficios.

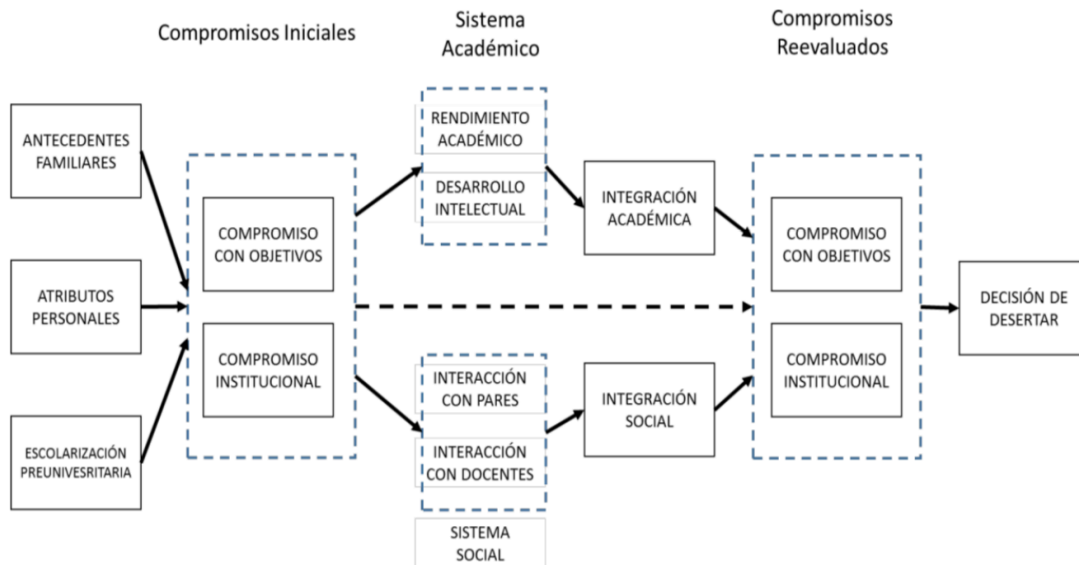


Ilustración 5: Modelo planteado por Tinto. Fuente: Tinto & Cullen, 1975

### 2.2.3 1985: Bean y su modelo basado en la productividad del ambiente laboral

En los siguientes años, Bean complementa los modelos desarrollados por Tinto y Spady a través de dos trabajos publicados en el 1980 y 1985. En el primero trabajo Bean toma los factores planteados por Tinto y Spady y define variables para testear si estos modelos tienen evidencia empírica. Cinco años más tarde, en el segundo trabajo amplía su estudio incluyendo en el análisis a estudiantes no tradicionales como respuesta a un cambio en el acceso a la educación superior de la época que hasta entonces era restringida solamente a un grupo de elite. Este cambio generó un aumento en la heterogeneidad del cuerpo estudiantil en la educación superior, generando un nuevo desafío de evidencia empírica de los modelos de Tinto y Spady (Bean & Metzner, 1985; Bean, 1980).

En artículo publicado en 1980, Bean planteó un conjunto de variables que podrían definir la deserción. Adicionalmente, con el objetivo de identificar las causalidades entre estas variables, define que las variables relacionadas con el antecedente del estudiante, tales como estado socioeconómico, desempeño académico previo y residencia actual (supuestamente diferente previo ingreso al instituto educacional), impactarían en los determinantes organizacionales o bien, de la institución educacional y estos a su vez, en la decisión de desertar (Bean, 1980). El conjunto de variables y sus relaciones se encuentran en la Tabla 1 e Ilustración 6

Tabla 1: Lista de variables planteadas por Bean. Fuente: Bean 1980

Variable	Definición
<i>Variables de Antecedentes</i>	
Desempeño Previo	Grado en que el estudiante ha demostrado sus logros académicos previos.

Estatus Socioeconómico	Grado en que los padres del estudiante han logrado estatus a través de la ocupación familiar.
Residente en el Estado	Si el estudiante es un residente del estado en donde la institución educativa está.
Distancia a Casa	Distancia de su residencia actual a la casa de sus padres
Tamaño de la Ciudad	Tamaño de la comunidad donde el estudiante pasó la mayor parte de su tiempo en su crecimiento.
<b><i>Determinantes Organizacionales (Basado en Price, 1977)</i></b>	
Rutina	Grado en que el rol de ser un estudiante es visto como una rutina.
Desarrollo	Grado en que un estudiante cree que el o ella se está desarrollando como resultado de ir a un Institución de Educación Superior (IES).
Valor práctico	Grado en que un estudiante percibe que su educación será utilizada para emplearse.
Calidad Institucional	Grado en que la IES es percibida como una proveedora de buena educación.
Integración	Grado en que el estudiante participa en relaciones primarias o cuasiprimarias (tiene amigos cercanos)
Promedio de Notas Universitario	Grado en que un estudiante demuestra su capacidad para desempeñarse en una IES.
Compromiso de Metas	Grado en que obtener un grado universitario es percibido como importante.
Comunicación	Grado en que la información sobre ser un estudiante es vista o entregada.
Justicia Distributiva	Grado en que un estudiante cree que es tratado justamente por la institución. Por ejemplo: recibe premios y castigos proporcionalmente a su esfuerzo realizado en su rol como estudiante.
Centralización	Grado en que un estudiante es tratado justamente por la Institución, Por ejemplo: Centros de estudiantes, consejeros, etc.
Advisor	Grado en que un estudiante cree que su advisor es útil.
Relación con Funcionarios	Nivel de contactos informales con los miembros de la Facultad.
Trabajo en el Campus	Necesidad de tener un trabajo en el campus universitario para permanecer en la escuela.
Área	El área de uno de los campos de estudio
Certeza	Grado en que un estudiante es poco indeciso en que se está especializando.

Alojamiento	Cuando una persona vive On Campus.
Organización del Campus	El número de miembros en la organización del campus
Oportunismo (Transferencia/ Trabajo /Hogar)	Grado en que un rol alternativo (como estudiantes, empleado o dependiente en casa de los padres) existe en el ambiente externo (otra universidad, en una empresa o volver a casa de los padres)
<b>Variables de Intervención</b>	
Satisfacción	Grado en que siendo un estudiante es visto positivamente
Compromiso Institucional	Grado de lealtad hacia la pertenencia del estudiante en la organización

Se destacan del listado las variables de calidad de la institución y la satisfacción del estudiante con la institución. Estos se pueden relacionar directamente con la satisfacción y su compromiso institucional, mientras que de manera transitiva con la decisión de desertar. Guiado por la literatura, Bean relacionó las variables y realizó regresiones lineales para el caso de los estudiantes mujeres y hombres. La relación entre estas variables se presenta a continuación en la Ilustración 6.

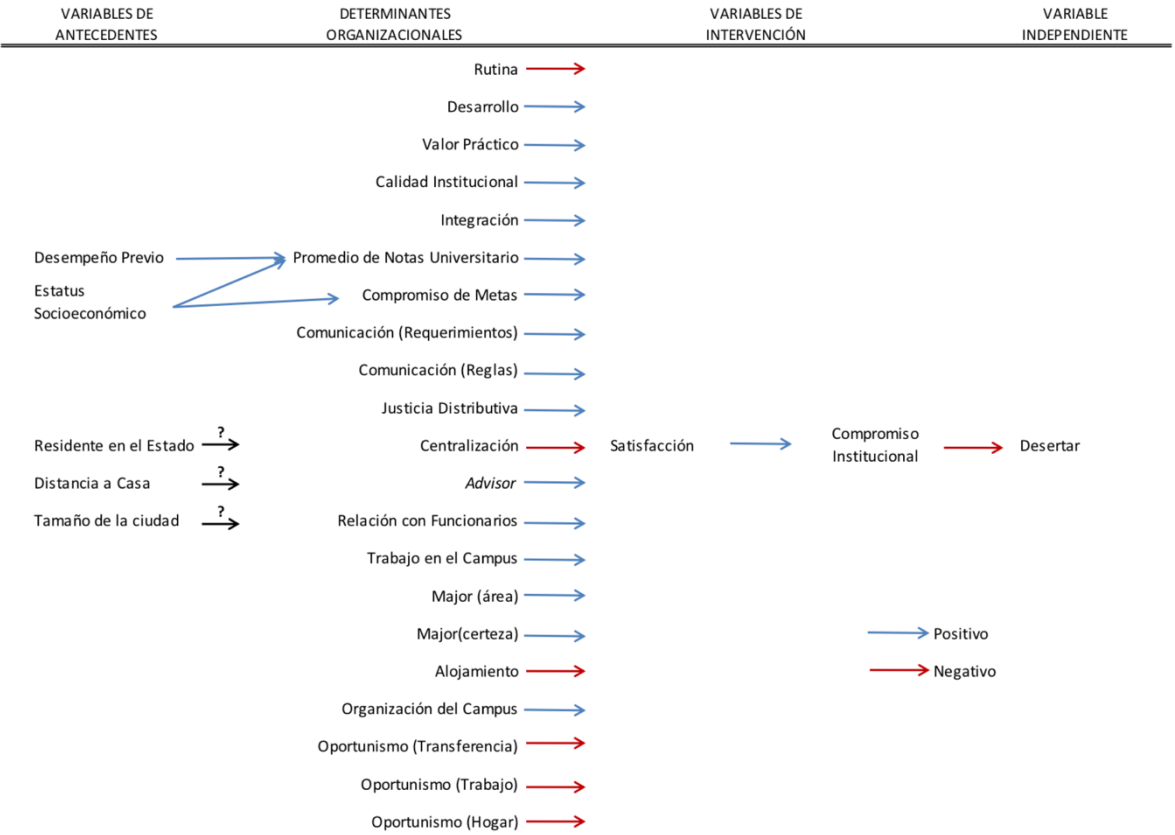


Ilustración 6: Relación entre variables planteadas por Bean. Fuente: Bean 1980



- **Comprensión de los Datos:** El entendimiento de los datos surge en forma iterativa desde el proceso de entendimiento del negocio y debe considerar al menos los siguientes aspectos:
  - a) Identificar las diversas fuentes de datos e información relacionadas con el problema u oportunidad declarada.
  - b) Recolectar datos e información y almacenar temporalmente en estructuras de datos que permitan su manejo posterior con las tecnologías que se dispongan para este efecto.
  - c) Explorar los datos recolectados y revisar estos en cuanto a su validez y confiabilidad para representar hechos o situaciones relacionadas directa o indirectamente con el problema u oportunidad declarada.
  - d) Comprender los datos y su relación con el problema u oportunidad definida.
  
- **Preparación de los Datos:** La preparación de los datos es una pieza clave para la fase de modelamiento y debe ser formalmente considerada para asegurar la obtención de un modelo acertado en relación con el error esperado por la empresa. Las principales etapas son las siguientes:
  - a) Seleccionar el dataset que será utilizado para el modelamiento.
  - b) Limpiar los datos para eliminar el ruido que pueda distorsionar el funcionamiento de los modelos.
  - c) Construir, integrar y formatear los datos, dejándolos preparados para la fase de modelamiento. Esta etapa genera varias transformaciones sobre los datos.
  
- **Modelamiento:** Esta fase es iterativa con la fase de preparación de los datos, puesto que en la medida que se van desarrollando los modelos, se requieren de nuevas integraciones de datos, mejoras en la construcción de los datos, nuevamente limpieza de los mismos, dado que la experiencia que se va desarrollando con cada iteración y con cada modelo, permite visualizar mejoras a ser aplicadas. Esta fase debe considerar al menos los siguientes aspectos:
  - a) Seleccionar los primeros modelos a utilizar (supervisados o no supervisados), de acuerdo a la estructura de los datos y conocimiento de los hechos o situaciones que éstos tienen en relación al problema definido. Establecer la medida del error o rendimiento que se aceptará para los modelos.
  - b) Modelar y aplicar los modelos, obteniendo y analizando los resultados de los mismos, realizando una comparación y estableciendo un ranking entre ellos. Seleccionar el o los mejores modelos de acuerdo a los criterios definidos.
  - c) Iterar si es necesario en el modelamiento o con la fase anterior de preparación de los datos.
  - d) Esta fase termina con la evaluación retrospectiva, si es posible de realizar, o cualquier otro mecanismo para evaluar la calidad del modelo realizado.
  
- **Evaluación:** Esta fase de evaluación consta de la generación de un piloto que tomen las recomendaciones que entregan los modelos, y que permitan, a través de la práctica, generar un

resultado comprobable que de cuenta cerca o lejos está el modelo de los resultados esperados en relación a los resultados del piloto. Es necesario realizar las siguientes actividades:

- a) Aprobar los modelos que finalmente entreguen los resultados esperados y documentarlos
  - b) Revisar el proceso realizado, comprendiendo y evaluando las etapas desarrolladas, con el objetivo de mejorar al siguiente ciclo de aplicación de la metodología.
  - c) Determinar los pasos siguientes, que podrían llevarnos a la puesta en funcionamiento o retroceder a la fase inicial de entendimiento, con el objetivo de mejorar los datos obtenidos e incluso, comprender en mayor forma el problema en estudio.
- **Implementación:** Esta fase se activa siempre y cuando existan modelos que puedan ser implementados, para lo cual debe realizarse lo siguiente:
- a) Documentar el procedimiento de aplicación del modelo en las situaciones reales del negocio.
  - b) Generar una estructura de casos para realizar una verificación respecto de la aplicabilidad del modelo a través del tiempo, identificando instancias de monitoreo necesarias y mantenimiento de los modelos requeridos.
  - c) Elaborar un plan de puesta en funcionamiento y documentación de las lógicas que serán ingresadas en los sistemas de información a objeto de mejorar su proceder y/o toma de decisiones.
  - d) Preparar las presentaciones finales y difusión respecto de los resultados obtenidos en el o los casos pilotos y en su aplicación en los procesos de manera cotidiana.

### **2.3.2 Minería de Datos**

En esta fase se realiza el modelamiento propiamente tal a través de la aplicación distintas técnicas. El objetivo es extraer patrones y conocimientos previamente desconocidos, los cuales se obtendrán a través del procesamiento de los datos.

La Minería de Datos comienza con la obtención de la base de datos y luego con el tratamiento de ésta. Este procesamiento se hace en base a métodos y modelos computacionales que hacen uso de técnicas estadísticas, inteligencia artificial, aprendizaje de máquina, entre otras, donde posteriormente, luego de reiteradas iteraciones y evaluaciones, se obtienen patrones ocultos en los datos que ayudan al conocimiento del negocio.

La Minería de Datos se utiliza para una vasta serie de tareas como se puede observar en la Ilustración 8, dentro de las cuales, los más usados dentro de las organizaciones son: Clasificación y Clusterización.





Ilustración 8: Aplicaciones de la Minería de Datos.  
Fuente: Han, Kamber & Pei 2012

La Clasificación es el proceso de encontrar una función o patrón que permita distinguir (predecir) clases entre los datos. Este proceso se realiza a partir de una base de datos de entrenamiento, la cual posee etiquetadas las clases respectivas a predecir. Con este entrenamiento, se obtiene un modelo parametrizado, para posteriormente predecir una base de datos de prueba que no posea etiquetadas las clases. Los modelos típicos de Clasificación son: Árboles de decisión, Redes Neuronales, Vectores de Soporte, entre otros.

La Clusterización, a diferencia de la clasificación, es un proceso que no entrena con un base de datos etiquetada, sino que crea clases (clústeres) en base al criterio de la máxima similaridad dentro de los clústeres y mínima similaridad entre clústeres. Con esto, cada clúster puede ser visto como una clase, donde se pueden etiquetar o nombrar en función de las características similares de las observaciones, con lo cual se puede conocer de mejor manera los datos y su entendimiento en términos del negocio. Los modelos típicos de clusterización son: K-means, Nearest-Neighbor Algorithm, DBSCAN, entre otros.

En la actualidad existen variadas técnicas de minería de datos que permiten la ejecución de las tareas anteriormente descritas. Respecto de la categorización y regresión es posible identificar máquinas de aprendizaje que, gracias a los avances tecnológicos, la implementación de estas se hace cada vez más accesible para los usuarios. Máquinas de aprendizaje tales como Support Vector Machine, Decision Tree y Artificial Neural Net han sido programadas y utilizadas como librerías en softwares de minería de datos. Para esta tesis se ha decidido aplicar las siguientes máquinas: (1) Support Vector Machine (SVM), (2) Decision Tree (DT), (3) Artificial Neural Net (ANN), (4) Logistic Regression (LR).

### 2.3.2.1 Máquinas de aprendizaje

#### 2.3.2.1.1 Support Vector Machine (SVM)

El Support Vector Machine (SVM) es un modelo del tipo de aprendizaje supervisado que utiliza algoritmos de clasificación y análisis de regresión. En términos prácticos, se puede explicar el funcionamiento del SVM como la clasificación de un conjunto de registros a través de la separación con un hiperplano, el cual minimiza el costo de error de clasificación de cada registro a una de las dos clases en estudio.

Formalmente se define el Support Vector Machine como la función de separación representada por un hiperplano en el espacio  $R^n$ , el cual maximiza el margen de separación. En una separación de puntos por dos clases, el problema se denomina del tipo linealmente separable. Bajo este concepto, en un espacio  $R^n$  con finitos puntos que representan observaciones, estos pueden ser separados por infinitos hiperplanos. Sin embargo, los algoritmos que aplican Support Vector Machine identifican la separación lineal óptima, el cual está definido como la máxima capacidad de generalización y el mínimo error empírico de clasificación. El margen de separación es definido como la distancia entre el par de paralelos canónicos, por lo que el margen es igual a dos veces el mínimo entre los puntos de entrenamiento y el hiperplano de separación. En estos puntos, los cuales minimizan la distancia de separación del hiperplano, son llamados vectores de soporte (support vectors). Sobre estos se obtienen las reglas de clasificación (Vercellis, 2009).

Matemáticamente, si se define  $\mathbf{w}$  como el vector de coeficientes del hiperplano y  $b$  como el intercepto, se define el hiperplano dado como:

$$\mathbf{w}'\mathbf{x} = b$$

Mientras que los dos hiperplanos paralelos canónicos son:

$$\mathbf{w}'\mathbf{x} - b - 1 = 0.$$

$$\mathbf{w}'\mathbf{x} - b + 1 = 0$$

Entonces el margen de separación  $\delta$  es definido como:

$$\delta = \frac{2}{\|\mathbf{w}\|}$$

Donde  $\|\mathbf{w}\| = \sum_{j \in N} w_j^2$ . En orden de determinar los coeficientes  $\mathbf{w}$  y el intercepto  $b$ , el hiperplano que optimiza la separación se define por la solución del siguiente problema cuadrático con restricciones lineales:

$$\begin{aligned} & \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s. a. } & y_i(\mathbf{w}'\mathbf{x}_i - b) \geq 1, \quad i \in \mathcal{M} \end{aligned}$$

Sin embargo, en la mayoría de los casos los  $m$  puntos no son linealmente separables. Entonces la ecuación anterior se formula de la siguiente manera:

$$\min_{w,b,e} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m e_i$$

$$s. a. \quad y_i(\mathbf{w}'\mathbf{x}_i - b) \geq 1 - e_i, \quad i \in \mathcal{M}$$

$$e_i \geq 0 \quad i \in \mathcal{M}$$

Donde  $C$  refleja el costo de error de clasificación.

A modo de ejemplo, en la Ilustración 9 se muestra la aplicación de SVM en donde divide un mapa de registros en dos categorías. Aquellos puntos ubicados al lado derecho del hiperplano son cateogrizados como no fuga (cuadrado azul) y aquellos ubicados al lado izquierdo como fuga (círculo rojo). El hiperplano generado estaría minimiza el costo de error de clasificación y adicionalmente, maximizaría la distancia entre los dos grupos dentro de un margen establecido.

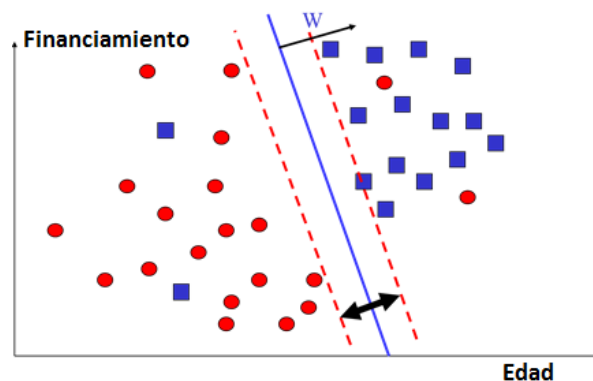


Ilustración 9: "Representación gráfica de la aplicación del algoritmo SVM"

## Parámetros

Los parámetros del Support Vector Machine son los costos asociados al error de clasificación. Por lo tanto, en esta tesis se optimizará el  $C$  que otorgue el mejor desempeño de los modelos.

### 2.3.2.1.2 Decision Trees (DT)

Los árboles de decisión, o Decision Tree (DT) en inglés, se basa en las teorías de decisiones para realizar clasificaciones a las bases de datos en donde se apliquen algoritmos minería de datos.

Los árboles de decisión segmentan los registros utilizando técnicas matemáticas y estadísticas, introduciendo el concepto de entropía (índice de incertidumbre o desorden), el cual sirve para identificar el siguiente atributo de segmentación.

Gráficamente los árboles se componen de nodos, ramas y hojas. Los nodos son puntos de unión, en donde se refleja una toma de decisión. Las ramas representan los arcos de conexión entre nodos, y las hojas son nodos terminales en donde se refleja la decisión final, es decir, la clasificación.

En el libro Data Science for Business(Provost & Fawcett, 2013) se muestran dos claros ejemplos de la segmentación y representación gráfica de un árbol de decisión. Respecto de la Ilustración 10, se puede ver la secuencia de nodos que reflejan la decisión de segmentación de un conjunto de datos. Los nodos finales reflejan la clasificación final con la probabilidad correspondiente.

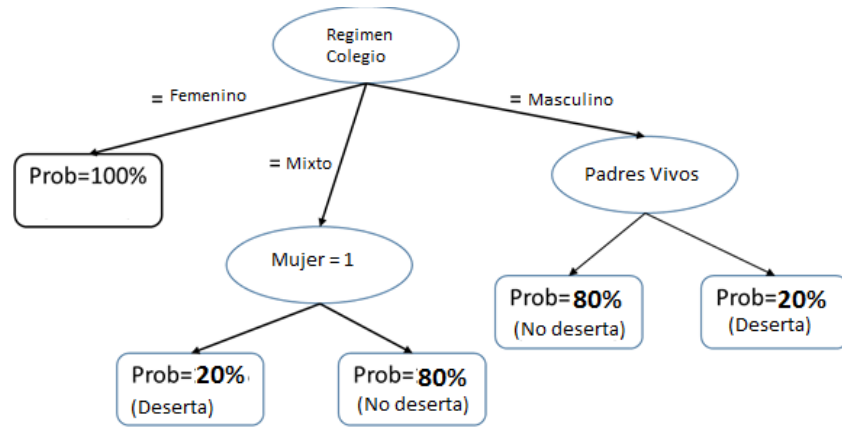


Ilustración 10: Representación gráfica del resultado obtenido por un algoritmo basado en la máquina de aprendizaje "Árbol de Decisiones"

En la Ilustración 11 se muestra de manera gráfica la segmentación realizada a través del árbol de decisión. Cada división refleja un nodo y los puntos dentro de un cuadro reflejan las hojas o nodos finales del árbol.

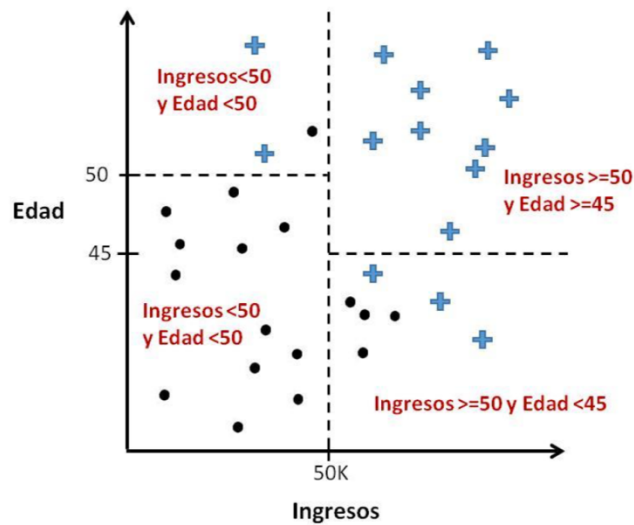


Ilustración 11: Representación gráfica de la segmentación, según algoritmo de Árbol de Decisión. Fuente (Provost & Fawcett, 2013)

La segmentación previamente descrita, requiere de algunos pasos específicos antes de implementar el algoritmo de clasificación. Los elementos necesarios a describir son:

## Reglas de Separación

Para cada nodo del árbol, se debe especificar el criterio por el cual se separarán los conjuntos de datos. Estas reglas varían por el número de nodos descendientes, el número de atributos y la evaluación de métricas. Dentro de estas últimas se encuentran dos mayormente usados: (1) Índice de Clasificación Errónea, (2) Entropía y (2) Gini.

El índice de Clasificación Errónea para un nodo  $q$  es calculado como:

$$Miscl(q) = 1 - \max_h p_h$$

Donde  $p_h$  es la proporción de observaciones en el nodo clasificados a la clase  $h$ . Este índice mide la proporción de ejemplos mal clasificados cuando todas las instancias del nodo  $q$  son asignados a la mayoría de la que ellos pertenecen, de acuerdo al principio de voto por mayoría (Vercellis, 2009).

Por otro lado, el índice de entropía para un nodo  $q$  es calculado como:

$$Entropy(q) = - \sum_{h=1}^H p_h \log_2 p_h$$

El índice en sí mismo es un índice de diversidad, ya que mide las diferencias de distribución de los grupos de clasificación.

Finalmente, el índice Gini para un nodo  $q$  es calculado como:

$$Gini(q) = 1 - \sum_{h=1}^H p_h^2$$

El índice Gini mide la uniformidad en la distribución de la clasificación de las instancias y sirve para distinguir la diferencia entre dos grupos categorizados de manera dicotómica.

## Criterios de Detención y Poda

En cada nodo del árbol se deben definir criterios de detención para el algoritmo. Se entiende como criterios de detención aquellos que indican si la construcción de una rama del árbol debería continuar recursivamente o bien, el nodo se debe considerar como hoja. Adicionalmente, con el objetivo de evitar el excesivo crecimiento de un árbol durante el desarrollo del algoritmo recursivo se generan criterios de pre-poda, como también reducir el número de nodos después de que el árbol haya sido generado (poda). En la literatura existen muchos criterios utilizados para la detención del algoritmo y poda, siendo los más comunes el tamaño del nodo, la pureza y el mejoramiento del rendimiento del árbol.

- **Tamaño del Nodo:** Este consistente en el número de observaciones bajo el nodo. El algoritmo recursivo terminará si el tamaño llega a ser menor de un umbral.

- **Pureza:** Se entiende como pureza la proporción de observaciones en un nodo que pertenecen a la misma clase. Mientras más alta es esta proporción, mayor pureza tiene el nodo. El algoritmo recursivo se detendrá una vez conseguido un mínimo nivel de pureza.
- **Mejoramiento:** Un algoritmo continuará la segmentación recursiva en una rama, si la división genera un mejoramiento en el desempeño del modelo. Para esta tesis la evaluación será a través de la precisión.

## Parámetros

Los parámetros para un árbol varían según el algoritmo utilizado. En el caso de esta tesis los parámetros utilizados para optimizar el rendimiento de los árboles serán: (1) Máximo Profundidad (Tamaño Poda), (2) Máximo tamaño separación, (3) Aplicación Poda y (4) Aplicación Prepoda.

### 2.3.2.1.3 Artificial Neural Network (ANN)

Las redes neuronales artificiales o artificial neural net (ANN) fueron introducidas inicialmente como concepto de red neuronal por los neurólogos (McCulloch & Pitts, 1943). Quince años más tarde, (Rosenblatt, 1958) generó el primer perceptrón simple basado en los conceptos de red neuronal, proponiendo así los fundamentos de una red neuronal artificial.

Una red neuronal se compone de redes de nodos llamados neuronas. Cada nodo recibe un conjunto de entradas provenientes de otros nodos y entregan una salida. Esta salida se compone por tres funciones:

- **Función de propagación:** Es la función que se compone por la sumatoria de las entradas multiplicados por un peso de interconexión.
- **Función de activación:** Es la función que modifica la función anterior (aprendizaje). Puede que la configuración de la red no tenga esta función, por lo que la salida es la misma función de propagación.
- **Función de transferencia:** Es la función que se aplica al valor que entrega la función de activación. Su principal es para acotar el rango de salida del nodo. Las más comunes son la función sigmoidea (intervalos entre 0 y 1) y la tangente hiperbólica (intervalos entre -1 y 1).

## Perceptrón

Gráficamente un perceptrón puede ser reflejado como la siguiente figura:

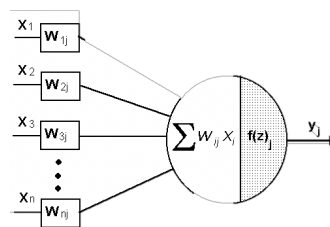


Ilustración 12: "Descripción gráfica de un perceptrón"

La Ilustración 12 es la forma más sencilla de una red neuronal y corresponde a una sola neurona de salida. Las entradas  $x_1, x_2, \dots, x_n$  son valores de entradas que se combinan con distintos pesos  $w_1, w_2, \dots, w_n$  y entregan un resultado de salida  $f(x)$ . Supongamos que los valores de los pesos ya han sido determinados durante el proceso de entrenamiento. Bajo este contexto, se puede determinar la predicción de una nueva observación bajo los siguientes pasos. Primero, la combinación lineal de los pesos con los valores de entrada (variables explicativas) para una nueva observación pueden ser calculadas como:

$$w_1x_1 + w_2x_2 + \dots + w_nx_n - E = w'x - \mathcal{E}$$

Entonces, la predicción de  $f(\mathbf{x})$  puede ser obtenida como:

$$f(\mathbf{x}) = g(w_1x_1 + w_2x_2 + \dots + w_nx_n - E) = g(w'x - \mathcal{E})$$

Donde  $g(\cdot)$  es la función de activación y su propósito es mapear la combinación lineal al conjunto de posibles valores de la variable dependiente. En un problema de clasificación binaria estos valores son definidos como  $[-1, 1]$ , entonces, una función de activación  $g(\cdot)$  para este caso sería la función signo ( $sgn(\cdot)$ ). Una vez identificada la función de activación, se implementa un algoritmo iterativo que determina los valores de los pesos  $w_i$ , examinando en secuencia, uno a uno las observaciones del vector  $x$ .

### Redes Multinivel de Prealimentación

Una estructura más compleja de las redes neuronales son las del tipo multinivel de prealimentación. Se compone de tres elementos principales:

- **Capa de Entrada:** Conjunto de neuronas que representan las variables/factores independientes.
- **Capa Oculta:** Conjunto de neuronas que reciben la información de las neuronas de entradas, estudia sus patrones y determina el peso de cada una de forma iterativa.
- **Capa de Salida:** Conjunto de neuronas que proporcionan la respuesta de la red neuronal. En un problema de clasificación, el número de neuronas en esta capa es igual al número de clases de clasificación.

A modo de ejemplo, la Ilustración 13 muestra la estructura básica de una red neuronal.

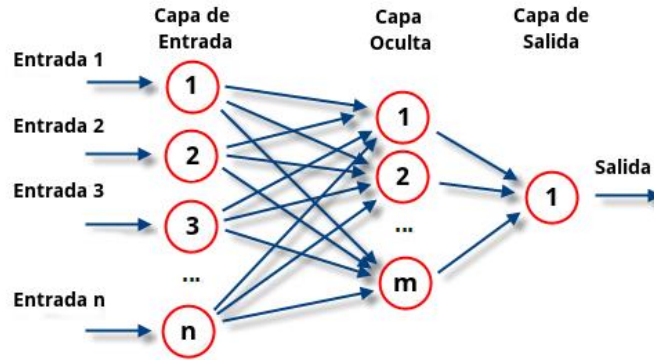


Ilustración 13: "Ejemplo de estructura de una red neuronal artificial"

Cada nodo opera básicamente como un perceptrón. En otras palabras, los pesos estarán asociados a cada arco conectado entre nodos, y cada nodo es asociado a un coeficiente de distorsión ( $E$ ) y una función de activación ( $g(\cdot)$ ). El método utilizado para determinar los pesos y los coeficientes de distorsión es denominado algoritmo de *backpropagation*, o en retropropagación en español. Un algoritmo de retropropagación inicia con valores de pesos  $w_i$  aleatorios. Posteriormente, cada instancia de la base de entrenamiento es examinada secuencialmente, generando una predicción para cada una y obteniendo desempeños de correctas y malas clasificaciones. Estos resultados son utilizados como retroalimentación para ajustar los pesos y realizar nuevamente la examinación con pesos y coeficiente de distorsión ajustados. Las redes neuronales tienen un alto desempeño predictivo, puesto que son capaces de capturar las relaciones complejas lineales y no lineales entre las variables independientes y dependientes. Sin embargo, necesita de grandes volúmenes de información para obtener un buen desempeño.

### Parámetros

Los algoritmos de redes neuronales varían en la configuración de los parámetros para su correcto funcionamiento. En el caso de esta tesis, se optimizarán dos parámetros principalmente: (1) Tamaño Capa Oculta y (2) Ciclos de Entrenamientos.

El Tamaño de la Capa Oculta indica la cantidad de nodos utilizados en esta capa de la red neuronal. Actualmente existen algoritmos que calculan de manera automática el tamaño ideal de la capa, siendo la mayoría de las veces  $n - 1$  nodos, donde  $n$  es la cantidad de variables dependientes. Sin embargo, no siempre este número entrega los mejores desempeños de las redes, por lo que en esta tesis la asignación automática será comparada con asignaciones manuales del número de nodos.

Los ciclos de entrenamientos hacen referencia al número de iteraciones que el algoritmo realizará para ajustar los valores de los pesos  $w_i$  y coeficiente de distorsión. Mientras mayor número de ciclos asignados, mayor probabilidad de obtener un mejor desempeño de las redes.



### 2.3.2.1.4 Logistic Regression (LR)

La Regresión Logística o *Logistic Regression* (LR) en inglés, es un caso especial de regresiones cuyo uso es para predecir el resultado de una variable dependiente categórica. Tiene bastante uso en los cálculos de probabilidades, donde se predice la ocurrencia de un evento en función de otros factores.

A modo de ejemplo, supongamos que la variable de respuesta  $y$  toma valores 0 y 1. De acuerdo a lo que postula la regresión logística, la probabilidad posterior  $P(y|x)$  de respuesta a la variable condicionada del vector  $x$  sigue una función logística definida como:

$$P(y = 1|x) = \frac{1}{1 + e^{-w^T x}}$$
$$P(y = 0|x) = \frac{e^{-w^T x}}{1 + e^{-w^T x}}$$

En las ecuaciones previamente mostradas, el algoritmo identifica los coeficientes  $w$  de forma iterativa, usualmente a través del método de máxima verosimilitud.

En general los modelos y algoritmos de regresión logística presentan la misma dificultad que las regresiones lineales, en otras palabras, pueden adolecer de problemas de multicolinealidad y sesgo.

#### Parámetros

En la etapa de entrenamiento, los algoritmos de regresiones logísticas obtienen las predicciones para un conjunto de observaciones  $X$  y al igual que los *support vector machines*, es posible calcular el costo de error de las clasificaciones erróneas. Adicionalmente, los algoritmos iteran con el objetivo de identificar el mejor conjunto de los valores  $w_i$ , en donde el costo total de clasificaciones erróneas busca ser minimizado. En esta tesis, el parámetro de las regresiones logísticas será el costo  $C$  para cada error de clasificación.

### 2.3.2.2 Interpretación y evaluación

En esta etapa se evalúa el desempeño de los modelos aplicados en la etapa anterior. En esta etapa el juicio experto juega un rol fundamental, ya que el investigador deberá evaluar si los patrones extraídos tienen sentido en el contexto que fueron aplicados. Cabe destacar que, en esta etapa, al igual que las anteriores, existe la posibilidad que se decida volver al primero paso o a una etapa previa según corresponda.

La literatura ha planteado distintas métricas para medir el desempeño predictivo de los modelos. Los más comunes son el Error de Clasificación y la Precisión de la Predicción (Accuracy), sin embargo, estas métricas miden el desempeño general de los modelos, asumiendo que todos los tipos de errores tienen el mismo costo, lo que no siempre es así en un contexto organizacional.

### Matriz de Confusión

La matriz de confusión es una tabla compuesta mayoritariamente por dos filas y dos columnas, las cuales contienen información sobre el desempeño de las clasificaciones predichas por un modelo de clasificación. Usualmente, las filas representan las instancias que el modelo predijo, mientras que las columnas las instancias observadas reales.

En el caso de una clasificación dicotómica, es decir, clasificación en dos clases, se generan dos tipos de clases nombradas positivas y negativas. Para cada observación se realizan predicciones de ambas clases, a través de la implementación de un algoritmo de técnica de minería de datos, y estas se comparan con el valor real de la clase.

Aquellas observaciones en que se predijo como clase positiva y efectivamente era de esa clase, son denominadas como True Positives (TP), o Verdadero Positivo en español; mientras que, si no lo eran, se evalúan como False Positive (FP), o Falso Positivo. Ocurre lo mismo para las clases negativas, asignando como True Negative (TN), o Verdadero Negativo, las predichas como negativas y efectivamente lo eran; mientras que False Negative (FN), o Falso Negativo, las con predicción de clase negativa pero efectivamente positivas (Shmueli, Patel, & Bruce, 2011).

		Clase Verdadera	
		+	-
Predicción	+	True Positive	False Positive
Clases	-	False Negative	True Negative

Ilustración 14: Ejemplo de Matriz de Confusión

Los datos que representa la matriz se describen a continuación:

- **True Positive o Verdaderos positivos:** Corresponden a los valores predichos como Positivos por el modelo (clase 1) de clasificación y que efectivamente corresponden a un valor Positivo, para los datos de prueba.
- **True Negative o Verdaderos negativos:** Son los valores predichos como Negativos por el modelo (clase 0) y que corresponden a un valor Negativo en el conjunto de datos.
- **False Negative o Falsos negativos:** Corresponden a valores predichos por el modelo como Negativos de forma incorrecta, ya que en el conjunto de datos corresponden a un valor Positivo.
- **False Positive o Falsos positivos:** Son los valores predichos como Positivos por el modelo, pero que en los datos de prueba corresponden a valores Negativos.

### Costos de Clasificación: Tipos de Errores I y II

A partir de la información entregada por la Matriz de Confusión, se definen los errores de Tipo I y II

		Clase Verdadera	
		+	-
Predicción	+	True Positive	Error Tipo I
Clases	-	Error Tipo II	True Negative

Ilustración 15: Ejemplo Error de Clasificación

Todo modelo predictivo debe contar con una tasa de error, se espera que sea así, de otra forma lo más probable es que se haya producido un sobreajuste.

Un factor importante a la hora de evaluar un modelo y compararlo con otros, reside en qué tipo de error es más tolerable; esto dependerá en gran medida del contexto para el que estemos generando un modelo predictivo.

Por ejemplo, predecir que un estudiante NO DESERTA cuando finalmente DESERTA es mucho más costoso que predecir que DESERTA cuando NO DESERTA (Conceptos Error Tipo I y Error Tipo II). De esta manera, la evaluación de los desempeños de cada modelo tendrá considerado la importancia de cada una de las clases.

### 2.3.2.3 Clasificadores

Los desempeños de los modelos pueden variar por distintos factores, tales como las variables escogidas o al manejo que se hizo de los datos, por lo tanto, se hace necesario contar con indicadores que permitan medir de manera objetiva su rendimiento.

Existen variados clasificadores al momento de evaluar un modelo, a continuación, se realizará una descripción de los más importantes.

- **Accuracy:** Fracción total de instancias correctamente clasificadas.

$$Accuracy: (TP+TN)/(TP+TN+FP+FN)$$

- **Precision:** Fracción de instancias clasificadas como positivas que en realidad son positivas.

$$Precision: TP/(TP+FP)$$

- **Recall:** También conocida como sensibilidad, corresponde a la fracción de las instancias positivas que fueron predichas positivas.

$$Recall: TP/(TP+FN)$$

- **Fscore o Fmeasure:** Corresponde a la media armónica entre precisión y recall.

$$Fscore: 2*(Recall*Precision)/ (Recall + Precision)/$$

- **Specificity:** Corresponde a la probabilidad de obtener un resultado negativo cuando la instancia efectivamente fue predicha como negativa.

$$Specificity: TN/(TN+FP)$$

- **Sensitivity:** Corresponde a la probabilidad de obtener un resultado positivo cuando la instancia efectivamente fue predicha como positiva.

$$Sensitivity: TP/(TP+FN)$$

- **Classification Error:** Corresponde a la suma de los Falsos Negativos más los Falsos Positivos dividido por el total de datos con los que se está trabajando

$$Classification Error: (FN+FP)/n$$

## CAPÍTULO 3: PLANTEAMIENTO ESTRATÉGICO Y MODELO DE NEGOCIOS

### 3.1 POSICIONAMIENTO ESTRATÉGICO

El posicionamiento estratégico del proyecto se basa en el modelo Delta, (Hax 2010). Este modelo sitúa al cliente en el centro de la estrategia, estableciendo cuáles son las opciones disponibles para generar un vínculo con el cliente proponiendo como ligar la estrategia y la ejecución a través del alineamiento de los procesos.

La vinculación con el cliente es la fuerza impulsora de la estrategia, ya que, atendiendo al cliente en forma distintiva, es posible atraerlo, satisfacerlo y retenerlo. Por medio del conocimiento de éste, es posible adecuar la oferta a través de la generación de nuevos productos.

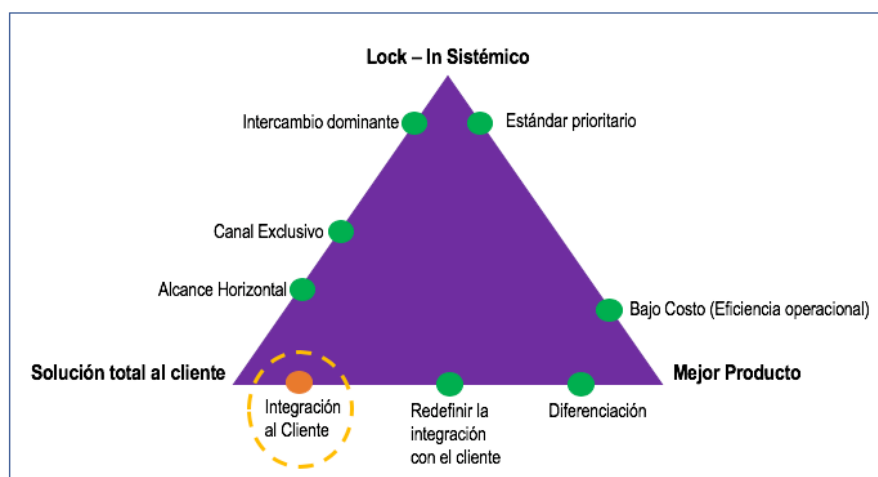


Ilustración 16: Posicionamiento Estratégico. Fuente: Elaboración Propia

El modelo que resulta del proyecto permite clasificar a los estudiantes recién ingresados de acuerdo a si continuarán o desartarán por algún motivo, basándose en la data histórica y así definir una estrategia de captación y retención hacia las necesidades especiales de cada estudiante.

### 3.2 MAPA ESTRATÉGICO

Para que el proyecto tenga una exitosa implementación se debe considerar el proceso de capacitación en la Unidad donde quedará operando este nuevo conocimiento.

Una vez que se haya realizado con éxito lo anterior se debe realizar la integración de los procesos y tecnología y así apoyar el proceso de gestión académica con lo cual se mantendrá una estrecha comunicación con los estudiantes de modo de aumentar la retención con el consiguiente de aumento de ingresos por concepto de arancel.

### 3.3 MODELO DE NEGOCIOS

El modelo de negocios, de acuerdo con la estructura propuesta por Johnson (2008), debe contener la propuesta de valor para el cliente, fórmula de utilidades, recursos claves y procesos claves.

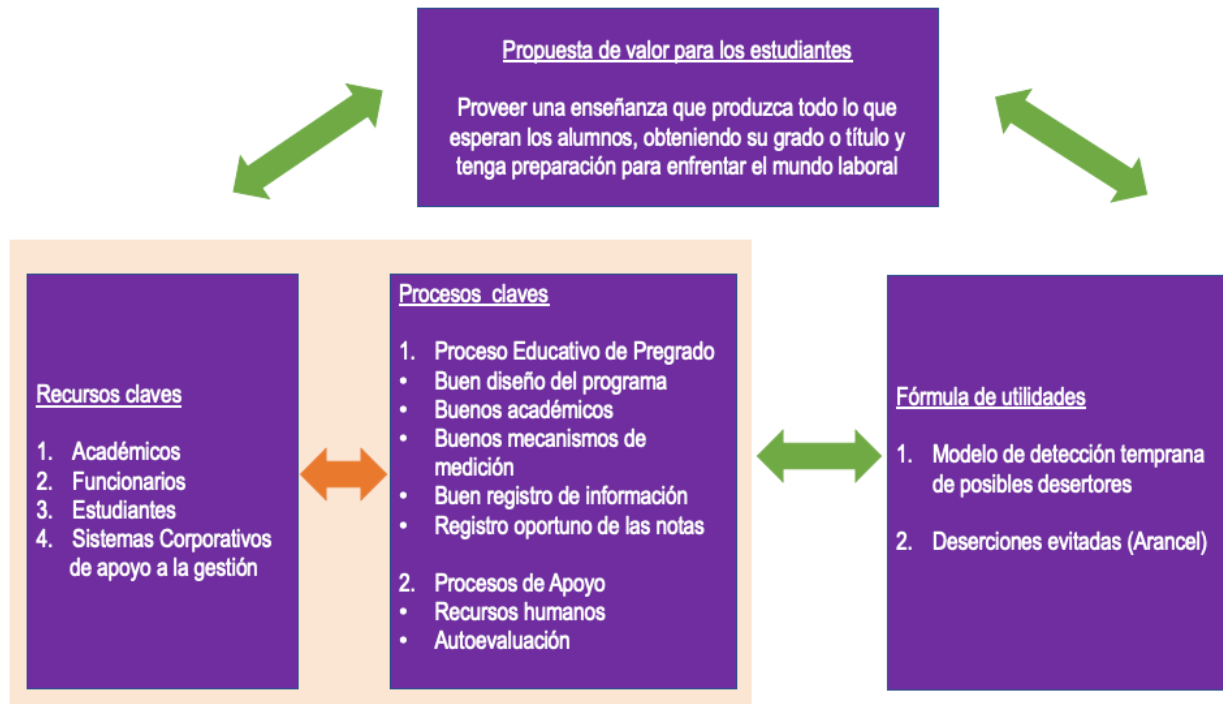


Ilustración 17: Modelo de Negocios. Fuente: Elaboración Propia con base en Johnson 2008

### 3.3.1 Propuesta de valor para el cliente

La propuesta de valor del proyecto para la Universidad es conocer los factores que influyen en la deserción estudiantil, pudiendo hacer las modificaciones en las mallas de las carreras o reforzamientos dirigidos a alumnos específicos o reestructurar la estructura de los contenidos de cada asignatura

### 3.3.2 Recursos claves

Los recursos claves son los académicos que son quienes imparten la docencia y son el vínculo de la entrega del conocimiento, los funcionarios quienes con su trabajo hacen que la Universidad marche adecuadamente, los estudiantes quienes son los clientes, pagan por un producto y servicio y financian la universidad, los sistemas corporativos de apoyo a la gestión académica que permiten registrar la información y obtener conocimiento de los datos.

### 3.3.3 Procesos claves

Los procesos claves son el proceso educativo de pregrado el cual debe tener un buen diseño del programa, considerando: horas cronológicas adecuadas en los mejores horarios, una distribución de la malla para balancear la carga académica, buenos docentes, los cuales cada vez son requeridos con grados más altos, si antes era requisito mínimo poseer un magíster, hoy se exige tener un grado de doctor, además existe la instancia de la evaluación docente, donde se puede ver como los estudiantes percibe el trabajo realizado por el académico.

Utilizando los recursos claves de sistemas corporativos y business report, se debe realizar el seguimiento de los procesos, al ser una tarea permanente, éste se transforma en un proceso en sí,

que cobra gran importancia para lograr la propuesta de valor controlando que se vaya ejecutando de acuerdo a lo planificado.

### **3.3.4 Fórmula de utilidades**

$$Utilidad = \delta * \alpha * (\gamma - \varphi)$$

*Donde*

$\delta$  = *Deserciones evitadas*

$\alpha$  = *Arancel del Programa de Estudios*

$\gamma$  = *Duración del programa de estudios*

$\varphi$  = *Años cursados en el programa de estudios*

La fórmula de utilidades se puede expresar de la siguiente manera, un estudiante que según el modelo indique que va a desertar, se le aplica reforzamiento y se evita que el estudiante deserte, esto haría sumar 1 el valor de  $\delta$ , ya que sería una deserción evitada, luego ésta se multiplica por el arancel del programa de estudios que está matriculado el estudiante ( $\alpha$ ), todo esto se debe multiplicar por la diferencia entre la duración del programa de estudios ( $\gamma$ ) y los años cursados que lleva el estudiante ( $\varphi$ ), esto último se explicaría como la permanencia futura del estudiante.

En el capítulo de financiamiento se explica con mayor detalle la fórmula de utilizadas.

### **3.3.5 Amenazas de nuevos competidores**

En el ámbito de la educación superior, si bien existe una amplia oferta educacional que no logra satisfacer la demanda, las barreras de entradas son muy altas, lo que implica que la amenaza de que surjan nuevos competidores sea muy baja.

### **3.3.6 Amenazas de productos y servicios sustitutos**

La amenaza de productos y servicios sustitutos es baja, podrían considerarse los programas online y algunos Institutos.

### **3.3.7 Poder de Negociación de clientes**

El poder de negociación de los clientes (estudiantes) es media, ya que estos han tomado mucha fuerza en los últimos años y al organizarse son capaces de producir cambios, sin embargo, se deben atener a los estatutos universitarios.

### **3.3.8 Poder de Negociación de Proveedores**

Cada vez entran más académicos al sistema, obteniendo grados a muy baja edad, subiendo los estándares de la universidad, esto genera una competitividad cada vez más grande lo que permite aún que la negociación de los proveedores (académicos) sea baja.

### **3.3.9 Rivalidad entre Competidores**

La rivalidad entre competidores es alta, ya que existen Universidades del CRUCH que han hecho que la Universidad deba invertir en difusión y competir tramo a tramo con ellas.

### **3.4 ANÁLISIS FODA**

#### **3.4.1 Fortalezas**

- Alta competencia del cuerpo académico.
- Alta calidad de oferta académica de Pregrado y Postgrado.

#### **3.4.2 Debilidades**

- Falta de comunicación entre las diferentes Unidades Académicas.

#### **3.4.3 Oportunidades**

- Concursos de financiamiento.

#### **3.4.4 Amenazas**

- Gran crecimiento del sistema universitario privado.
- Burocratización creciente del sistema de financiamiento y compras, comparado con las universidades privadas.



## CAPÍTULO 4: ANÁLISIS DE LA SITUACIÓN ACTUAL

La arquitectura de procesos actual de la Universidad de ACME, utilizando la metodología de ingeniería de negocios propuesta por Barros (2004), permite adecuar los patrones de negocios desde los 4 macroprocesos hasta el proceso específico de interés del proyecto.

### 4.1 ARQUITECTURA DE PROCESOS

La arquitectura de procesos actual de la Universidad, utilizando la metodología de ingeniería de negocios propuesta por Barros (2004), permite adecuar los patrones de negocios desde los 4 macroprocesos hasta el proceso específico de interés del proyecto.

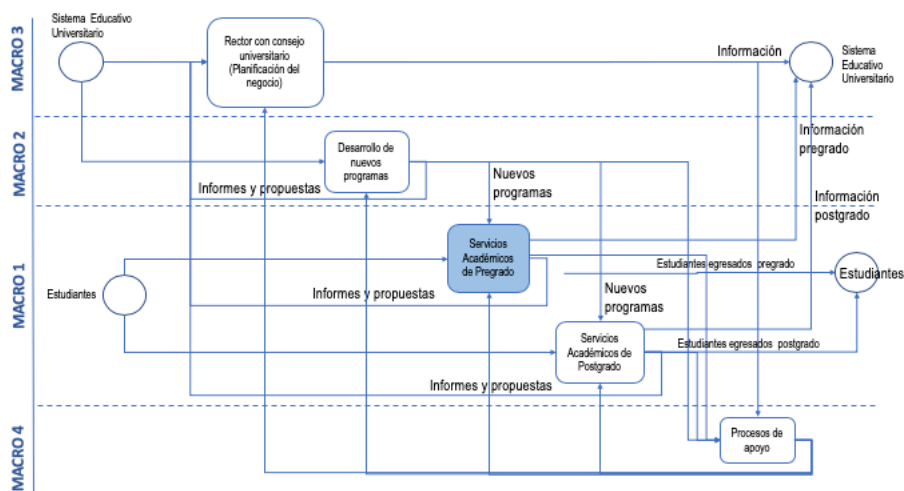


Ilustración 18: Macroprocesos Universidad. Fuente: Elaboración Propia

#### 4.1.1 Macroproceso 1: Cadena de Valor.

Procesos que entregan el servicio de educación al estudiante, desde la postulación y admisión, pasando por matrícula, inscripción de asignaturas y terminando con la graduación.

#### 4.1.2 Macroproceso 2: Desarrollo de nuevas capacidades.

Desarrollo de nuevos programas, los nuevos programas de pregrado responden a necesidades de la institución y es una decisión de alto nivel donde el rector está involucrado, por otro lado, los programas de postgrado se centran en una necesidad de las Unidades Académicas las cuales formulan el nuevo programa y deben contar con la supervisión y aprobación del Departamento de Postgrado y Postítulo.

#### 4.1.3 Macroproceso 3: Planificación del Negocio.

Planificación de la Universidad, contempla todos los procesos y actividades de planificación que realizan los altos directivos, entiéndase Rector, Prorector, Vicerrectores, Consejo Universitario y Decanos, quienes en conjunto con sus respectivos equipos bajan líneas de planificación hasta el más bajo nivel directivo.

#### 4.1.4 Macroproceso 4: Procesos de Apoyo.

Procesos de Apoyo a los otros macroprocesos, contempla Aranceles, Recursos Humanos, Matrícula, donde intervienen funcionarios del área TI, de las unidades académicas y de los Departamentos de Pregrado y Postgrado y Postítulo.

#### 4.1.5 Servicios Académicos de Pregrado

Los servicios académicos de pregrado, se preocupan tanto de la administración de la relación con el estudiante, la administración de la relación con el académico y la gestión de los procesos académicos. Es aquí donde surge la necesidad de incorporar un proceso que identifique a los alumnos que tienen la probabilidad de desertar una vez que comienza el año académico.

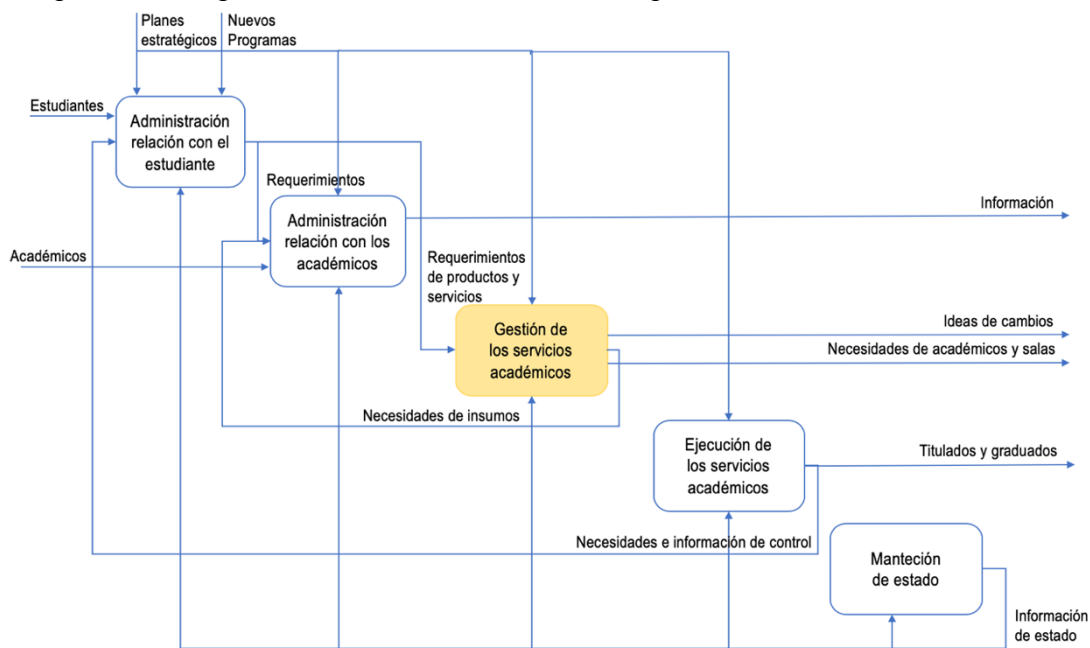


Ilustración 19: Servicios Académicos de Pregrado. Fuente: Elaboración Propia

##### 4.1.5.1 Administración de relación con el estudiante

La administración de la relación con el estudiante se realiza principalmente en las unidades académicas, pero en caso de que éstas no puedan resolver alguna situación es escalada a nivel del Departamento de Pregrado, por ejemplo, solicitudes de reincorporación, problemas atendibles, entre otros.

##### 4.1.5.2 Administración de relación con el académico

La administración de la relación con los académicos se realiza principalmente en los departamentos de las unidades académicas, contrataciones, asignación de cursos, evaluación académica, entre otros.

##### 4.1.5.3 Gestión de los Servicios Académicos

La gestión de servicios académicos de pregrado, se preocupan tanto de la implementación de nuevos programas, de la planificación y control de procesos académicos y debe decidir la entrega de los procesos académicos.

#### 4.1.5.4 Ejecución de los Servicios Académicos

Consiste en la ejecución de los planes diseñados en “Gestión de los Servicios Académicos”.

#### 4.1.5.5 Mantenimiento de Estado

Consiste en el registro de la información de cada una de las actividades que componen los procesos y el flujo que existe en cada uno de ellos para apoyar el proceso de toma de decisiones.

#### 4.1.6 Gestión de los Servicios Académicos

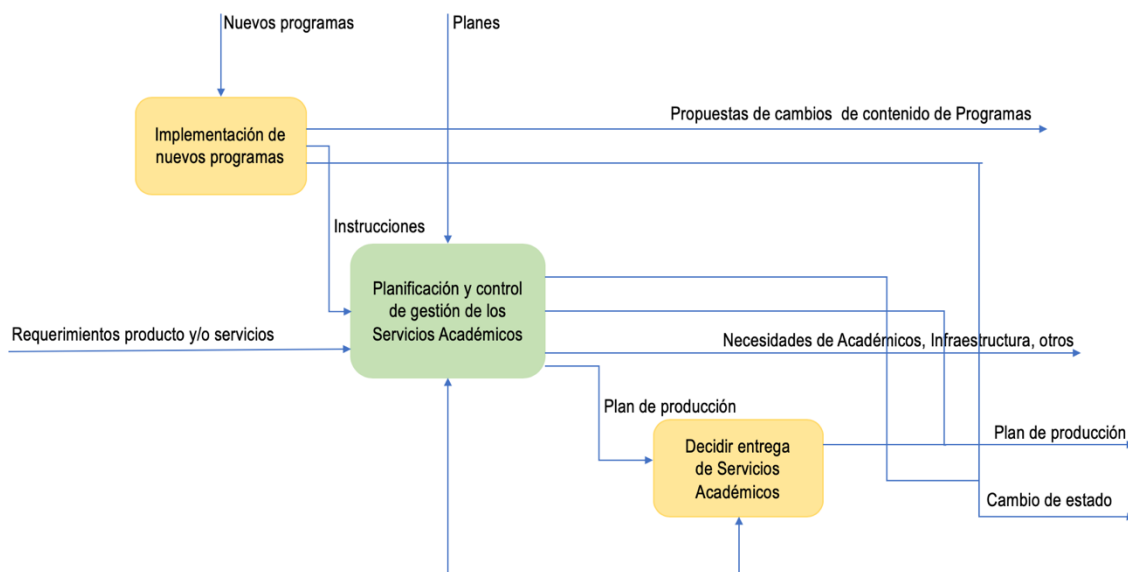


Ilustración 20: Gestión de Servicios Académicos. Fuente: Elaboración Propia

##### 4.1.6.1 Implementación de nuevos programas

El Departamento de Pregrado solicita la creación del programa de Alertas Tempranas y Seguimiento a Estudiantes, definiendo nuevas normativas dirigidas a los docentes ya que son ellos quienes deben ingresar las notas y la asistencia de los alumnos de sus cursos.

##### 4.1.6.2 Planificación y control de servicios académicos

Aquí es donde el proyecto deberá ejecutarse, ya que, hasta el momento, no se encuentra desarrollado el seguimiento de los estudiantes con posibilidad de desertar según su rendimiento académico.

##### 4.1.6.3 Decidir entrega de procesos académicos

Dada la planificación del punto anterior, en esta actividad se debe decidir cuándo se hará la intervención en la nivelación a los estudiantes identificados como posibles desertores y planes de acción orientados a entregar el apoyo que sea necesario.

## 4.2 CUANTIFICACIÓN DEL PROBLEMA U OPORTUNIDAD

### 4.2.1 Antecedentes: Deserción en la Educación Superior

La tasa de Retención de estudiantes en Educación Superior (ES), en especial la de primer año, es uno de los indicadores más utilizados a nivel internacional para evaluar la eficiencia interna de las instituciones terciarias, considerando que la mayor deserción de estudiantes se da en ese período. Desde el año 2007, el Servicio de Información de Educación Superior (SIES), del Ministerio de Educación, calcula la tasa de Retención de 1er año para programas de Pregrado, actualizándola año a año para ver su evolución según las distintas variables del sistema.

Esta Tasa de Retención de 1er año se calcula como el cociente entre el número de estudiantes que ingresan como alumnos de primer año a una carrera o programa en un año determinado, y el número de esos mismos estudiantes que se mantienen como estudiantes antiguos en la misma institución al año siguiente, expresado en términos porcentuales<sup>3</sup>.

A modo de comparación, se presentan dos tasas alternativas y relacionadas con la Retención de 1er año: Tasa de Persistencia de 1er año en la misma institución y Tasa de Persistencia de 1er año en Educación Superior. En todas estas tasas se está usando como referencia el Pregrado.

La Tasa de Persistencia de 1er año en la misma institución de Educación Superior considera el cociente entre el número de estudiantes que ingresan como alumnos de primer año a una carrera o programa en un año determinado, y el número de esos estudiantes que se mantienen en la misma institución al año siguiente, expresada en términos porcentuales<sup>4</sup>.

En este caso, se consideran como persistentes tanto a aquellos estudiantes que se mantienen en la cohorte e institución del año anterior, como a los que reingresaron a la misma institución como estudiantes de 1er año en algún programa.

La Tasa de Persistencia de 1er año en Educación Superior (ES) considera el cociente entre el número de estudiantes que ingresan como alumnos de primer año a una carrera o programa en un año determinado, y el número de esos mismos estudiantes que se mantienen en alguna institución de Educación Superior al año siguiente, expresada en términos porcentuales. En este caso, se consideran como persistentes en ES a aquellos estudiantes que aparecen vinculados a una institución de educación terciaria, independiente de la cohorte e institución de origen.

La tasa de Retención de 1er año de Pregrado para la cohorte 2017, que considera la continuidad de la cohorte de origen, es de 74,0% considerando al total de instituciones de educación superior, siendo levemente más baja que la tasa de Persistencia en la misma institución que alcanza el 75,2%. Por su parte, la tasa de Persistencia en Educación Superior alcanza el 81,9%.

---

<sup>3</sup> Estos antecedentes corresponden a un resumen de los principales hallazgos de la Retención de 1º año de carreras y programas que SES publica en tablas con mayor detalle y desagregación en un archivo Excel disponible en [www.mifuturo.cl](http://www.mifuturo.cl)

<sup>4</sup> Fuente: [www.mifuturo.cl](http://www.mifuturo.cl)

Tipo de institución	Retención 1 <sup>er</sup> año	Persistencia 1 <sup>er</sup> año en la misma IES	Persistencia 1 <sup>er</sup> año en ES	N° casos Retención
				1 <sup>er</sup> año
CFT	68,7%	69,9%	75,2%	57.826
IP	70,9%	71,3%	76,4%	116.143
Universidades	78,7%	80,7%	89,2%	140.641
<b>Total general</b>	<b>74,0%</b>	<b>75,2%</b>	<b>81,9%</b>	<b>314.610</b>

Ilustración 21: Tasas de retención, persistencia en la misma institución y en Educación Superior de 1º año para carreras de Pregrado por tipo de institución, cohorte 2017. Fuente: www.mifuturo.cl

A su vez, es posible observar que la distancia entre la tasa de Retención de 1er año y la de Persistencia en la misma institución es mayor en universidades (2,0 puntos porcentuales, en adelante p.p.) que en IP y CFT (0,4 y 1,2 p.p. respectivamente). Lo anterior implica que es más frecuente que estudiantes reingresan o se cambien de carrera (partiendo como estudiante nuevo) en la misma universidad a que ello ocurra en IP e CFT.

Por otra parte, la tasa de Persistencia de 1er año en Educación Superior incluye a los estudiantes retenidos en la misma cohorte e institución, a quienes persisten en la misma institución con independencia de la cohorte de origen, y agrega a los estudiantes que se cambian a otra institución distinta a la de la cohorte de origen. En este caso, a nivel general, la tasa de Persistencia de 1er año en Educación Superior es de 81,9%. En el caso de las universidades la tasa es de 89,2%, en los IP 76,4 y en CFT 75,2%.

Al analizar la Retención de 1er año y Persistencia en ES, se observa que mientras en las universidades la diferencia entre ambas tasas es de 10,5 p.p., en los IP y CFT es la mitad (5,5 y 6,5 p.p. respectivamente). Lo anterior, da cuenta de la mayor frecuencia con que los estudiantes de universidades se mantienen en el sistema, aunque no continúen en la misma cohorte e institución de origen, respecto de los estudiantes de CFT e IP.

#### 4.2.2 Retención de 1er año

En adelante, el presente informe se centrará en la tasa de Retención de 1er año, que es aquella que sigue la historia de la cohorte.

A nivel general, la tasa de Retención de 1er año para programas regulares de Pregrado para la cohorte de estudiantes 2017 es de 74,0%, subiendo 1,6 p.p. respecto del año anterior, y manteniendo la tendencia a un alza moderada registrada sucesivamente desde la cohorte del año 2011 (donde la tasa fue de 68,5%).

Tipo de institución	2013	2014	2015	2016	2017
CFT	63,9%	64,5%	65,7%	66,7%	68,7%
IP	66,1%	67,3%	67,6%	68,5%	70,9%
Universidades	75,0%	76,3%	76,9%	77,9%	78,7%
<b>Total general</b>	<b>69,5%</b>	<b>70,5%</b>	<b>71,2%</b>	<b>72,4%</b>	<b>74,0%</b>

Ilustración 22: Evolución de Retención de 1º año por tipo de Institución. Fuente: www.mifuturo.cl

Las universidades tienen mayores tasas de Retención de 1er año que los IP y los CFT. Para la cohorte del año 2017, las universidades alcanzan una tasa de Retención de 1er año de 78,7%. En IP y CFT la retención de 1er año es de 70,9% y 68,7% respectivamente.

En los tres tipos de institución se observa un aumento sucesivo de esta tasa, respecto a la cohorte 2013, subiendo 4,8 p.p. en IP y en CFT, y 3,7 p.p. en universidades.

Las Carreras Profesionales (con y sin licenciatura) tienen tasas de Retención de 1er año mayores que las Carreras Técnicas para la cohorte 2017. En el caso de las Carreras Profesionales alcanzan una tasa de Retención de 1er año de 78,0%, en tanto las Carreras Técnicas 69,3%. En la comparación de las cohortes 2013 y 2017, se observa que las Carreras Profesionales suben 4,3 p.p. y las Carreras Técnicas 4,7 p.p.

Tipo de carrera	2013	2014	2015	2016	2017
Carreras Profesionales	73,7%	75,0%	76,0%	76,6%	78,0%
Carreras Técnicas	64,6%	65,6%	66,1%	67,5%	69,3%
<b>Total general</b>	<b>69,5%</b>	<b>70,5%</b>	<b>71,2%</b>	<b>72,4%</b>	<b>74,0%</b>

Ilustración 23: Evolución de Retención de 1º año por tipo de carrera. Fuente: www.mifuturo.cl

En cuanto al tipo de carrera y jornada, se observa que la Retención de 1er año es mayor entre los matriculados en carreras en jornada diurna que en jornada vespertina. Para la cohorte 2017, la tasa de Retención de 1er año diurna es de 78,7% y la vespertina 64,5%. Al considerar solo carreras profesionales, la tasa de Retención de 1er año es de 81,1% en jornada diurna y 63,5% en vespertina.

Al considerar Carreras Profesionales y áreas del conocimiento, se observa que las áreas de Salud y Educación poseen mayor Retención de 1er año para la cohorte 2017, con 82,7% y 82,2% respectivamente. Por su parte, las áreas con menor retención son Ciencias Básicas (71,9%) y Humanidades (74,9%).

Entre las Carreras Técnicas, las áreas con mayor tasa de Retención de 1er año son Ciencias Básicas (74,5%) y Salud (72,5%). Las áreas con menor retención son Humanidades (61,7%) y Derecho (63,0%). Cabe destacar que el área de Tecnología representa casi el 40% de la matrícula en este tipo de carreras, y posee una tasa de retención de 1er año de 66,1%.

Al considerar la acreditación institucional, se observa que las instituciones de Educación Superior acreditadas tienen persistentemente mayor tasa de Retención de 1er año que aquellas que no lo están.

Para la cohorte del año 2017, las instituciones acreditadas muestran una tasa Retención de 1er año de 75,7% y las sin acreditación de 57,9%.

Igual tendencia se observa al analizar la acreditación según el tipo de institución. En el caso de las universidades, para el año 2017, las acreditadas presentan una tasa de Retención de 1er año de 79,7% y las no acreditadas de 61,1%.

En los IP, las acreditadas poseen una tasa de 73,3% y las no acreditadas de 56,6%. En los CFT, las instituciones acreditadas tienen una tasa de Retención de 1er año de 69,9% y las no acreditadas 56,9%.

Acreditación institucional (año decohorte)	2013	2014	2015	2016	2017
<b>CFT</b>	63,9%	64,5%	65,7%	66,7%	68,7%
Acreditada	66,1%	65,9%	66,2%	67,1%	69,9%
No acreditada	54,3%	53,9%	59,8%	61,1%	56,9%
<b>IP</b>	66,1%	67,3%	67,6%	68,5%	70,9%
Acreditada	66,9%	69,2%	69,2%	70,0%	73,3%
No acreditada	56,9%	51,3%	51,9%	58,0%	56,6%
<b>Universidades</b>	75,0%	76,3%	76,9%	77,9%	78,7%
Acreditada	76,0%	78,4%	78,5%	79,0%	79,7%
No acreditada	60,2%	57,0%	61,1%	56,3%	61,1%
<b>Total general</b>	69,5%	70,5%	71,2%	72,4%	74,0%

Ilustración 24: Evolución de Retención de 1º año por tipo de Institución y condición de acreditación Institucional. Fuente: www.mifuturo.cl

El panorama regional muestra que para la cohorte 2017, las regiones que lideran la Retención de 1er año entre las Carreras Profesionales son las regiones del centro sur del país: Maule (82,3%), Los Ríos (81,3%), La Araucanía (80,8%) y Biobío (79,8%), a las que se suma la Región de Arica y Parinacota con 79,3%. Por el contrario, menores tasas de retención en carreras profesionales se dan en las regiones del norte, tales como Tarapacá (70,5%), Antofagasta (71,7%) y Atacama (73,1%), junto con la región de O'Higgins (74,2%).

En cuanto a las Carreras Técnicas, las regiones con más altas tasas de Retención de 1er año para la cohorte 2017 son Los Ríos (73,3%), Maule (73,0%), Arica y Parinacota (72,9%) y O'Higgins (72,8%), mientras las menores tasas se observan en Atacama (65,1%), Tarapacá (65,5%) y Antofagasta (66,6%).

### 4.2.3 Deserción en la Universidad ACME

A continuación, se presenta un análisis estadístico de la deserción y una caracterización de los alumnos que hacen abandono de sus estudios.

Tabla 2: Total de estudiantes que desertan por año de ingreso. Fuente: Elaboración Propia

Total de estudiantes que desertan por año de ingreso			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
2010	1155	319	1474
2011	1359	370	1729
2012	1159	409	1568
2013	1310	413	1723
2014	1524	332	1856
2015	1705	373	2078
2016	1717	526	2243
<b>Total general</b>	<b>9929</b>	<b>2742</b>	<b>12671</b>

En la Tabla 2, se observa que desde el 2014, hay una tendencia al alza a nivel Universidad, ya que 332 alumnos se retiraron el primer año, 373 el 2015 y 526 el 2016. Según la información proporcionada en la Tabla 3, durante este periodo, la mayor cantidad de alumnos que desertaron fueron hombres.

Tabla 3: Total de estudiantes que desertan por género. Fuente: Elaboración Propia

Total de estudiantes que desertan por género			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
FEMENINO	3792	868	4660
MASCULINO	6137	1874	8011
<b>Total general</b>	<b>9929</b>	<b>2742</b>	<b>12671</b>

De un 100% de los alumnos que desertan, el 60,43% proviene de establecimientos particulares subvencionados (dependencia 3), seguido por 28,63% que proviene de establecimientos municipales (dependencias 1, 2 y 5). Si sumamos ambas cifras, se obtiene un 89,06%, por lo tanto, es posible concluir que ambos grupos son representativos al momento de desertar.

Tabla 4: Estudiantes que desertan según tipo de dependencia. Fuente: Elaboración Propia

Estudiantes que desertan según Tipo de dependencia			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
MUNICIPAL (DEPENDENCIA 1, 2, 5)	31,75%	28,63%	31,07%
NULL	3,81%	5,98%	4,28%
PARTICULAR PAGADO (DEPENDENCIA 4)	3,34%	4,96%	3,69%
PARTICULAR SUBVENCIONADO (DEPENDENCIA 3)	61,10%	60,43%	60,96%
<b>Total general</b>	<b>100,00%</b>	<b>100,00%</b>	<b>100,00%</b>

El promedio de notas de enseñanza media (NEM) de los alumnos que abandonan sus estudios es de 5,5, lo cual genera un puntaje de 524,9 puntos al momento de presentarse a rendir la Prueba de Selección Universitaria (PSU).



Tabla 5: Estudiantes que desertan según promedio de puntajes. Fuente: Elaboración Propia

Estudiantes que desertan según Promedio de Puntajes			
	Etiquetas de columna	0	1 Total general
Promedio de PROMEDIO_NOTAS_EM		5,6	5,5 5,6

	Etiquetas de columna	0	1 Total general
Promedio de PUNTAJE_NEM		539,2	524,9 536,2

Respecto a los puntajes obtenidos en la Prueba de Selección Universitaria, se observa que, en las pruebas de Lenguaje y Matemáticas, en promedio, bordean los 550 puntos, sin embargo, los puntajes obtenidos en Ciencias e Historia, son 364,4 y 307,5 para los alumnos que desertan, respectivamente

Tabla 6: Estudiantes que desertan según promedio de puntajes. Fuente: Elaboración Propia

Estudiantes que desertan según Promedio de Puntajes			
	Etiquetas de columna	0	1 Total general
Promedio de PUNTAJE LENGUAJE		556,2	553,1 555,5

	Etiquetas de columna	0	1 Total general
Promedio de PUNTAJE_MATEMATICA		571,4	555,5 568,1

	Etiquetas de columna	0	1 Total general
Promedio de PUNTAJE_HISTORIA		292,4	307,5 295,6

	Etiquetas de columna	0	1 Total general
Promedio de PUNTAJE_Ciencias		385,7	364,4 381,1

Se observa también que un 61,52% de los alumnos que desertan, han ingresado a la Universidad, inmediatamente después de salir del colegio. Este análisis se respalda con la Tabla que se muestra a continuación.

Tabla 7: Estudiantes que desertan según proceso al cual postula. Fuente: Elaboración Propia

Estudiantes que desertan según Proceso con el cual postulan			
Cuenta de ID	Etiquetas de columna	0	1 Total general
Etiquetas de fila			
NULL		28,38%	30,23% 28,78%
PROCESO ACTUAL		64,48%	61,52% 63,84%
PROCESO ANTERIOR		7,14%	8,24% 7,38%
Total general		100,00%	100,00% 100,00%

La información entregada en la Tabla 7, respalda la que se muestra a continuación, ya que los alumnos de 18 años son los que presentan un mayor porcentaje de deserción, seguido por los alumnos de 19 años. Ambos grupos suman un 67,14%.

Tabla 8: Total de estudiantes que desertan por edad. Fuente: Elaboración propia

Total de estudiantes que desertan por edad			
Cuenta de ID	Etiquetas de columna	0	1 Total general
Etiquetas de fila			
18		41,77%	40,37%
19		26,83%	26,77%
20		9,47%	11,60%
21		5,14%	5,73%
17		4,19%	4,12%
22		3,53%	3,32%
23		2,76%	1,46%
24		1,95%	1,39%
25		1,18%	1,02%
26		0,90%	1,02%
27		0,73%	0,69%
28		0,41%	0,66%
29		0,22%	0,44%
30		0,13%	0,33%

Haciendo un análisis respecto a la relación que existe entre la preferencia del alumno por la casa de estudios al momento de postular se obtiene lo siguiente:

Tabla 9: Estudiantes que desertan según preferencia

Estudiantes que desertan según Preferencia			
Cuenta de ID	Etiquetas de columna	0	1 Total general
Etiquetas de fila			
1		42,31%	40,44%
2		14,46%	13,86%
3		7,36%	6,67%
4		3,65%	3,14%
5		1,66%	2,30%
6		0,77%	1,28%
7		0,41%	0,40%
8		0,18%	0,33%
9		0,11%	0,33%
10		0,08%	0,15%
NULL		29,01%	31,11%
<b>Total general</b>		<b>100,00%</b>	<b>100,00%</b>

Un 40,44% de los alumnos que desertan, postularon a esta Universidad como primera opción, seguido por 13,86% que la tenían como segunda en su lista de prioridades.

Tabla 10: Estudiantes que desertan respecto a tipo de Institución anterior. Fuente: Elaboración Propia

Estudiantes que desertan respecto a tipo de institucion anterior			
Cuenta de ID	Etiquetas de columna	0	1 Total general
Etiquetas de fila			
☐ Centro de Formación Técnica		0,57%	0,88%
☐ Instituto Profesional		1,42%	1,90%
☐ NULL		83,51%	86,54%
☐ Universidad		14,49%	10,69%
<b>Total general</b>		<b>100,00%</b>	<b>100,00%</b>

Del total de alumnos que desertan, el 10,69% estudió anteriormente en la Universidad, 1,90% en Institutos Profesionales y sólo 0,88% en Centros de Formación técnica. Sin embargo, hay un 86,5 de datos nulos, es decir, no se tiene información.

El 93,25% de los alumnos que desertan, son solteros y en general, no trabajan durante el periodo académico, ya que según la Tabla 12, el 79,36% de un 100% no tienen trabajo.

Tabla 11: Estudiantes que desertan según estado civil. Fuente: Elaboración Propia

Estudiantes que desertan según estado civil			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
0	0,02%	0,00%	0,02%
1	95,53%	93,25%	95,04%
2	0,44%	0,66%	0,49%
3	0,22%	0,11%	0,20%
4	0,02%	0,11%	0,04%
NULL	3,77%	5,87%	4,22%
<b>Total general</b>	<b>100,00%</b>	<b>100,00%</b>	<b>100,00%</b>

0	Sin datos
1	Soltero
2	Casado
3	Separado
4	Viudo

Tabla 12: Estudiantes que desertan respecto a la posesión de trabajo. Fuente: Elaboración Propia

Estudiantes que desertan respecto a posesión de trabajo			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
NO	84,02%	79,36%	83,01%
NULL	3,77%	5,87%	4,22%
OCASIONALMENTE	5,17%	6,60%	5,48%
SI, PERMANENTEMENTE	4,34%	4,63%	4,40%
SIN DATOS	2,71%	3,54%	2,89%
<b>Total general</b>	<b>100,00%</b>	<b>100,00%</b>	<b>100,00%</b>

Del 100% de los alumnos que desertan, el 29,43% provienen de familias que tienen ingresos brutos mensuales desde \$172.001 hasta los \$262.000, seguido por un 21,33% que posee ingresos entre los \$262.001 hasta los \$345.000.

Tabla 13: Estudiantes que desertan según ingreso bruto total del grupo familiar. Fuente: Elaboración Propia

Estudiantes que desertan según ingreso bruto total del grupo familiar			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
2	30,53%	29,43%	30,29%
3	22,27%	21,33%	22,07%
4	12,00%	12,84%	12,18%
1	9,35%	8,86%	9,24%
5	7,95%	7,26%	7,80%
NULL	3,77%	5,87%	4,22%
6	4,23%	4,27%	4,24%
7	3,50%	3,61%	3,53%
12	2,42%	2,52%	2,44%
8	1,37%	1,57%	1,41%
9	1,00%	0,95%	0,99%
11	0,75%	0,84%	0,77%
10	0,87%	0,66%	0,82%
0	0,02%	0,00%	0,02%
<b>Total general</b>	<b>100,00%</b>	<b>100,00%</b>	<b>100,00%</b>

TRAMO	DESDE	HASTA
1	0	\$172.000
2	\$172.001	\$262.000
3	\$262.001	\$345.000
4	\$345.001	\$456.000
5	\$456.001	\$570.000
6	\$570.001	\$700.000
7	\$700.001	\$910.000
8	\$910.001	\$1.183.000
9	\$1.183.001	\$1.840.000
10	\$1.840.001	o más

Respecto a la cantidad de alumnos que desertan según la cantidad de personas que componen su grupo familiar, se observa que del 100%, un 28,41% de los alumnos, provienen de familias de 4 integrantes, seguido por un 19,73%, cuyo grupo familiar es de 5 personas.

Tabla 14: Estudiantes que desertan según número de personas que componen el grupo familiar. Fuente: Elaboración Propia

Estudiantes que desertan según número de personas que componen el grupo familiar			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
4	28,99%	28,41%	28,86%
5	21,71%	19,73%	21,28%
3	18,86%	17,80%	18,63%
6	9,89%	9,48%	9,80%
2	6,91%	6,46%	6,81%
NULL	3,77%	5,87%	4,22%
7	3,47%	4,74%	3,75%
0	2,55%	3,14%	2,68%
8	1,65%	2,30%	1,79%
9	0,92%	0,66%	0,86%
1	0,63%	0,51%	0,61%
10	0,33%	0,47%	0,36%
11	0,12%	0,18%	0,13%
14	0,04%	0,07%	0,05%
22	0,00%	0,04%	0,01%
12	0,08%	0,04%	0,07%
13	0,04%	0,04%	0,04%
16	0,01%	0,04%	0,02%
17	0,01%	0,04%	0,02%
18	0,01%	0,00%	0,01%
<b>Total general</b>	<b>100,00%</b>	<b>100,00%</b>	<b>100,00%</b>

En resumen, el perfil del alumno que deserta es el siguiente: el 60, 43% provienen de colegios particulares subvencionados, en promedio, las notas de enseñanza media es 5,5. En la PSU obtienen 553,1 en Lenguaje, 555,5 en Matemáticas, 307,5 en Historia y 364,4 en Ciencias. Ingresan inmediatamente después de salir del colegio. El 93,25% es soltero, el 79,36% no tiene trabajo remunerado y provienen de familias de escasos recursos.

Las familias están compuestas en su mayoría por 4 integrantes y tal como se muestra en la Tabla 15, un 54,70% de esas familias, cuenta sólo con una persona que genera ingresos para el hogar.

Tabla 15: Estudiantes que desertan según el n° de personas que trabajan en forma remunerada en el grupo familiar. Fuente: Elaboración Propia

Estudiantes que desertan según el n° de personas que trabajan en forma remunerada en el grupo familiar			
Cuenta de ID	Etiquetas de columna		
Etiquetas de fila	0	1	Total general
0	11,54%	11,74%	11,59%
1	58,26%	54,70%	57,49%
2	22,92%	23,81%	23,12%
3	2,80%	3,03%	2,85%
4	0,57%	0,62%	0,58%
5	0,11%	0,15%	0,12%
6	0,02%	0,07%	0,03%
NULL	3,77%	5,87%	4,22%
<b>Total general</b>	<b>100,00%</b>	<b>100,00%</b>	<b>100,00%</b>

A pesar de lo relevante y valiosa de esta información, no es posible utilizarla para apoyar el trabajo de esta Tesis ya que hay muchos datos faltantes, por lo tanto, se decidió analizar las notas de los alumnos para predecir la deserción.

## CAPÍTULO 5: PROPUESTA DE DISEÑO DE PROCESOS

Para lograr los objetivos del proyecto, es necesario realizar un diseño al interior de Planificación y Control de Servicios Académicos e implementar un sistema de seguimiento y monitoreo de las calificaciones de las asignaturas con el propósito de generar las alarmas correspondientes, evitando posibles deserciones.

### 5.1 DIRECCIONES DE CAMBIO Y ALCANCE

A continuación, se detallan las variables de cambio, según la metodología de Ingeniería de Negocios del Dr. Barros (2017) con el objetivo de dirigir los esfuerzos hacia el diseño propuesto.

a	Estructura de Empresa y Mercado	Situación Actual	Situación Propuesta
a.1	Servicio integral al cliente	No	Si
a.2	Lock-in sistémico	No	No
a.3	Integración con proveedores	No	No
a.4	Estructura interna: centralizada o descentralizada	Descentralizada	Centralizada
a.5	Toma de decisiones: centralizada o descentralizada	Descentralizada	Centralizada

Tabla 16: Variable "Estructura de Empresa y Mercado"

b	Anticipación	Situación Actual	Situación Propuesta
b.1	Planificación con centros de aprendizaje	No	Generar políticas internas que obligue a los docentes a ingresar las calificaciones de manera regular y periódica, con el objetivo de identificar los alumnos que deben nivelarse académicamente para poder retener.
b.2	Modelo predictivo de riesgo de deserción	No	Modelo basado en técnicas de business intelligence para identificar automáticamente a los alumnos con riesgo de deserción
b.3	Monitoreo de centros de aprendizajes	No	Revisión constante del ingreso de las calificaciones y asistencia a los sistemas de información

Tabla 17: Variable "Anticipación"

c	Coordinación	Situación Actual	Situación Propuesta
c.1	Reglas	No	Reglas de negocio automatizadas que genere dos tipos de alarmas. La primera cuando pasado el plazo establecido las notas y asistencia no hayan sido ingresados y la

			segunda, cuando los datos anteriores ingresen dentro de la categoría de riesgo.
c.2	Jerarquía	Es utilizada cuando los Secretarios de Estudios solicitan formalmente el ingreso de información a los sistemas de información corporativos	Elaborar reglas de negocio que identifique a los actores relevantes que es necesario notificar en caso que la información solicitada no esté disponible con el objetivo de informar a las autoridades y que éstos últimos sean quienes la pidan.
c.3	Colaboración	No	Generar talleres con los docentes para darles a conocer la importancia de lo que se les está solicitando y el impacto negativo que genera para la Universidad no ingresar esa información de manera oportuna
c.4	Partición	No	No

Tabla 18: Variable: "Coordinación"

d	Prácticas de trabajo	Situación Actual	Situación Propuesta
d.1	Lógica de negocio automatizada o semiautomatizada		
	Elaboración de informes de deserción	No	Lógica de negocio 100% automatizada para la confección de estadísticas en línea
	Revisión de calificaciones y asistencia en línea	No	Lógica de negocio 100% automatizada para la confección de estadísticas en línea
d.2	Lógica de apoyo a actividades tácitas	No	Elaboración y medición de indicadores de gestión
d.3	Procedimientos de comunicación e integración	No	Comunicación y Coordinación permanente entre las Unidades a cargo
d.4	Lógica y procedimientos de medición de desempeño y control	No	Elaboración y medición de indicadores de gestión

Tabla 19: Variable "Prácticas de Trabajo"

e	Integración de procesos conexos	Situación Actual	Situación Propuesta
e.1	Proceso aislado	No	Crear un proceso que se relacione con las Unidades correspondientes
e.2	Todos o la mayor parte de los procesos de un macroproceso	No	No

e.3	Dos o más macros que interactúan	No	No
-----	----------------------------------	----	----

Tabla 20: Variable "Integración de Procesos Conexos"

f	Mantenimiento Consolidada de Estado	Situación Actual	Situación Propuesta
f.1	Datos Propios	No	Crear un set de datos con las variables que se desean monitorear
f.2	Integración con datos de otros sistemas de empresa	No	Integrar con el sistema que tiene las notas y asistencia de los alumnos
f.3	Integración con datos de sistemas de otras empresas	No	No

Tabla 21: Variable "Mantenimiento Consolidada de Estado"

## 5.2 DISEÑO DETALLADO DE PROCESOS

El proyecto de diseño no afecta a la arquitectura de procesos de la Universidad ACME ya que solo considera el diseño de procesos dentro de la cadena de valor “Gestión de Servicios Académicos”, donde los procesos a diseñar se ubican dentro del proceso “Planificación y Control de Servicios Académicos”.

Al diseñar el nuevo proceso to be, el objetivo es asegurar que se cumplan los objetivos planteados en este proyecto de tesis, para lo cual es necesario incorporar dentro del proceso, todas las actividades detalladas, reglas de negocio, interacción con actores relevantes y productos a generar.

### 5.2.1 Diseño en IDEF0

La apertura de la cadena de valor de “Servicios Académicos de Apoyo”, no cambia de forma estructural, así como tampoco el Macroproceso de “Gestión de Servicios Académicos”. El proceso que se desea diseñar es “Planificación y Control de Gestión de los Servicios Académicos”, incorporando el desarrollo de modelos predictivos y el diseño de un programa de seguimiento y nivelación a los alumnos que estén en riesgo de deserción. El proceso se detalla a continuación

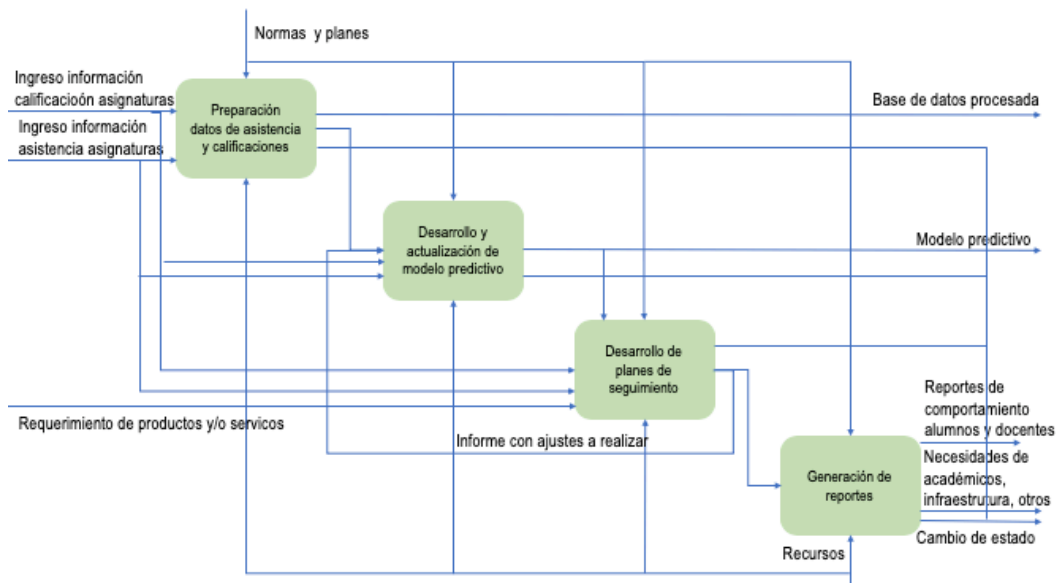


Ilustración 25: Planificación y control de servicios académicos. Fuente: Elaboración Propia

En la apertura de “Desarrollo y actualización de modelo predictivo” se encuentra en el seguimiento proactivo a los alumnos con probabilidad de desartar, el cual, identificará a los estudiantes que requieren nivelación o apoyo académico.

Este último punto se trabajará en detalle en el proceso de “Desarrollo de planes de seguimiento” el cual, enviará los resultados de los planes a “Generación de reportes” donde se evaluarán los indicadores de gestión y el impacto de los planes desarrollados.

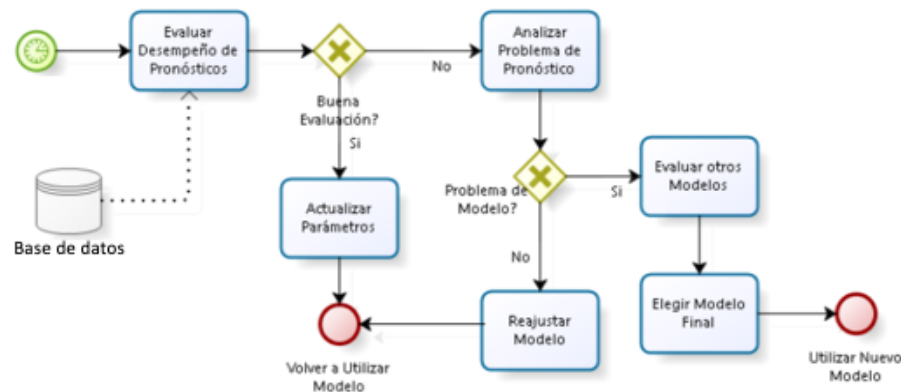


Ilustración 26: Desarrollo Modelo Predictivo

### 5.3 LÓGICA DE NEGOCIOS

El motor principal que mueve el diseño propuesto es la incorporación de inteligencia de negocios al proceso de “Gestión de Servicios Académicos”, en particular, en el área de “Planificación y Control



de Gestión de los Servicios Académicos”, por lo cual, se requiere una retroalimentación constante de las notas de las asignaturas y de la asistencia de los alumnos a clases por parte de los docentes.

Para poder minimizar el error que los modelos puedan entregar, dado que el supuesto con el cual que se está trabajando es que los alumnos desertan el primer año cuando tienen bajas calificaciones en uno más ramos y nula o poca asistencia, es que el comportamiento de los alumnos varía con el tiempo, por lo tanto, un modelo fijo no sería flexible a los cambios ni sería un modelo basado en la realidad. Para esto, la lógica de negocios está basada en una lógica iterativa de mejora continua como la propuesta por Deming la cual consiste en un ciclo llamado PDCA (Plan, Do, Check, Act) por sus siglas en inglés. En la Ilustración 26, las etapas instanciadas para el problema son: Modelar, Planificar, Analizar y Evaluar, las cuales son alimentadas por las fuentes de datos internas de la Universidad.

El **Modelar** consiste en la creación y/o modificación de los modelos de inteligencia utilizados para la creación de programas, siendo alimentado por las bases de datos internas. Esta etapa constituye todos los procesos enfocados al análisis y preparación de los datos, desarrollo y modificación de modelos y creación de listados de notas y asistencia de alumnos de cada curso con su probabilidad de deserción.

El **Planificar** es toda acción enfocada a la creación del programa de nivelación de alumnos y sus respectivas definiciones, respecto a como se va a realizar. Esta etapa finaliza cuando el apoyo al alumno es entregado.

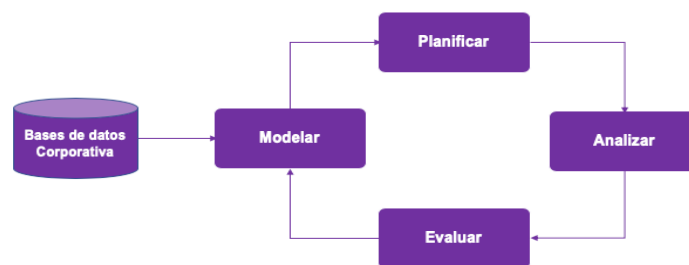


Ilustración 27: Lógica de Negocios Propuesta. Fuente: Elaboración Propia

**Analizar** es toda acción realizada por los tutores o profesores a cargo, donde estos, en base al programa de reforzamiento, realizan las respectivas clases de nivelación las cuales fueron planificadas previamente. Esta etapa termina cuando todas las clases de nivelación del programa finalizan, es decir, se tiene una evaluación del progreso del estudiante.

**Evaluar** es la etapa donde, en base a los resultados obtenidos por la etapa previa, se analiza la capacidad predictiva que tuvo el modelo y se evalúan diferentes indicadores para esto. Esta evaluación entrega indicios si el modelo predijo de manera aceptable o hay que realizar modificaciones para mejorar el error de predicción.

De esta manera la lógica se repite y se generan instancias de aprendizaje y retroalimentación las cuales permiten el tomar acciones de mejora dentro de las etapas y como ciclo general. Se espera que en la medida que haya iteraciones realizadas, los indicadores se estabilicen, dado que las iteraciones iniciales pueden generar cambios abruptos en los indicadores a evaluar.

#### **5.4 PRUEBA DE LA LÓGICA DE NEGOCIOS**

Inicialmente se entregaron dos grandes bases de datos. La primera base de datos contenía todos los datos de ingreso del estudiante, la cual es entregada por el DEMRE más información que es capturada por la Universidad en el momento que el alumno se matricula en una determinada carrera. La segunda base de datos contenía los datos “operacionales”, es decir, las notas de los alumnos en la carrera que ingresan.

En un comienzo la idea original era trabajar con ambas bases de datos y buscar patrones ocultos en los datos de ingreso de los estudiantes, con el objetivo que el apoyo se realice desde el comienzo por medio de la identificación “ex -antes” de los posibles alumnos desertores.

Al analizar y trabajar con los datos anteriormente mencionados, no se encontró ninguna correlación entre los datos que explique la deserción, básicamente porque el grupo de alumnos que ingresa tienen prácticamente las mismas características. La muestra es bastante homogénea, por lo cual, no fue posible encontrar factores que discriminen.

En reuniones con funcionarios del área de Estadísticas de la Universidad y docentes de algunas carreras, mencionaron que, en su experiencia, lo que influía en la deserción, son las notas de los alumnos, en particular, si las dos primeras calificaciones son deficientes (en determinados ramos), los alumnos se retiran y dejan de asistir.

Es por esta razón que se trabajó con los datos operacionales, es decir, las notas de los alumnos del primer semestre, con el fin de encontrar cuales son los patrones que definen la deserción.

Para la prueba de la lógica de negocios, se realizarán dos grandes iteraciones a los datos entregados, donde inicialmente se trabajó con las notas de los alumnos de primer año desde el 2010 al 2016 y en una segunda iteración se realizó por carrera. A continuación, se detallan las fases de la metodología utilizada y su aplicación en la organización.

##### **5.4.1 Entendimiento del negocio**

La fase de entendimiento del negocio se realizó en base al juicio experto de los académicos y funcionarios que trabajan en la Universidad, los cuales conocen la realidad académica, socioeconómica y personal de los alumnos. También conocen el funcionamiento administrativo y de gestión del ingreso de notas a los sistemas Corporativos, lo cual es responsabilidad de los docentes, sin embargo, esto último es un problema todos los semestres ya que no se cuenta con información oportuna.

#### **5.4.1.1 Determinar los objetivos del negocio**

El objetivo general que se destaca en este proyecto de tesis tiene directa relación con uno de los objetivos principales de la Organización, que es, “diseñar un proceso que genere alertas e identifique tempranamente a los alumnos de pregrado de primer año con mayor probabilidad de desertar, a través de la construcción de un modelo predictivo que permita detectar a los posibles desertores de pregrado, por medio de la caracterización de una serie de atributos personales, socioeconómicos, grupo familiar y de rendimiento universitario de primer año”.

Específicamente se quiere identificar los predictores y los valores que éstos toman con el propósito de poder intervenir oportunamente sobre los alumnos.

#### **5.4.1.2 Evaluación de la situación actual**

Actualmente, la Universidad, no cuenta con un proceso de Alertas Tempranas ni tampoco conoce las razones por las cuales los alumnos desertan. Esto último genera problemas en los procesos de Acreditación y también representa un problema de gestión interna ya que se desconoce las razones que lleva a los alumnos a abandonar sus estudios, impidiendo así el mejoramiento de sus procesos de gestión internos.

#### **5.4.1.3 Determinación de los objetivos de la Minería de Datos**

Los principales objetivos de la Minería de Datos son:

- Análisis de los Datos de la organización.
- Entendimiento del Comportamiento a través del análisis.
- Obtención de Modelos que posean un ajuste adecuado en términos de predicción.

#### **5.4.2 Entendimiento de los datos**

En esta etapa se seleccionan las variables y registros con los que finalmente se trabajará. De acuerdo a la literatura, existen técnicas estadísticas y matemáticas que se recomiendan para realizar esta selección, con el objetivo de identificar las variables que cumplan con los siguientes requisitos:

- Potencialmente explicativas del fenómeno en estudio
- Cuenten con poco o nulo error de registro
- Disponibilidad en el futuro en caso de realizar el análisis nuevamente.
- Medibles antes que ocurra el evento en estudio

#### **5.4.2.1 Recopilación de los datos iniciales**

Tal como se mencionó anteriormente los datos con los cuales se trabajó corresponden a las notas del periodo 2010 al 2016. La carrera con la cual se decidió trabajar corresponde a Diseño Industrial. Lo anterior corresponde a una decisión interna de la Universidad.

Los datos entregados corresponden a las notas de los alumnos de primer año a partir del 2010 más la información si deserta o no.

### 5.4.2.2 Descripción de los datos

Las variables que se presentan a continuación corresponden a los ramos de primer año de la carrera Diseño Industrial. Cada asignatura tiene controles de cátedra más talleres (según corresponda). La escala utilizada es de 1 a 7.

Nombre de la Variable	Descripción	Semestre	Tipo de Variable
DISI6011	Cultura y Diseño	1	Numérica
DISI6012	Forma y Estructura	1	Numérica
DISI6013	Representación Técnica Digital	1	Numérica
DISI6014	Dibujo I	1	Numérica
DISI6015	Taller de Diseño Industrial I	1	Numérica
HUMI6011	Inglés I	1	Numérica
PPSB0001	Taller de Comunicación Efectiva	1	Numérica
DISI6022	Tecnología de los Materiales I	2	Numérica
DISI6023	Representación Técnica Digital I	2	Numérica
DISI6024	Dibujo II	2	Numérica
DISI6025	Taller de Diseño Industrial II	2	Numérica
HUMI6021	Inglés II	2	Numérica
PPSB0002	Taller para el Desarrollo del Pensamiento Lógico Deductivo	2	Numérica

Tabla 22: Malla carrera Diseño Industrial. Fuente: Elaboración Propia

### 5.4.2.3 Exploración de los datos

Cada asignatura tiene 5 o 7 notas durante el semestre, sin embargo, en la tabla 9, es posible observar la cantidad de valores perdidos que existen para las tres primeras notas para cada ramo de la carrera, por lo anterior, se considerarán solo éstas, ya que las notas posteriores tenían más del 95% de registros faltantes.

Se tienen 415 registros de alumnos, que corresponden a los estudiantes desde el 2010 hasta el 2016 del primer semestre de la carrera.

Ramos	Missing values	% Missing values
DISI6011_NP1	208	50,24
DISI6011_NP2	208	50,24
DISI6011_NP3	209	50,48
DISI6012_NP1	212	51,21
DISI6012_NP2	216	52,17
DISI6012_NP3	216	52,17
DISI6013_NP1	211	50,97
DISI6013_NP2	212	51,21
DISI6013_NP3	367	88,65

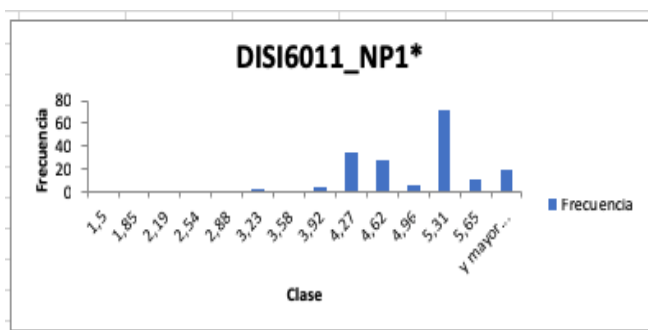
DISI6014_NP1	217	52,42
DISI6014_NP2	217	52,42
DISI6014_NP3	225	54,35
DISI6015_NP1	211	50,97
DISI6015_NP2	211	50,97
DISI6015_NP3	211	50,97
HUMI6011_NP1	242	58,45
HUMI6011_NP2	242	58,45
HUMI6011_NP3	370	92,03
PPSB0001_NP1	213	51,45
PPSB0001_NP2	215	51,93
PPSB0001_NP3	400	96,62

Tabla 23: Valores perdidos por asignatura, 1er semestre. Fuente: Elaboración Propia

A continuación, se presenta un análisis de los datos por cada asignatura. Se presenta la distribución de las notas de los alumnos para las tres primeras evaluaciones. En cada tabla se puede observar la cantidad de alumnos del total de la muestra.

### DISI6011: Cultura y Diseño

Clase	Frecuencia
1,5	1
1,85	0
2,19	0
2,54	1
2,88	1
3,23	3
3,58	2
3,92	5
4,27	35
4,62	28
4,96	6
5,31	72
5,65	11
y mayor...	19

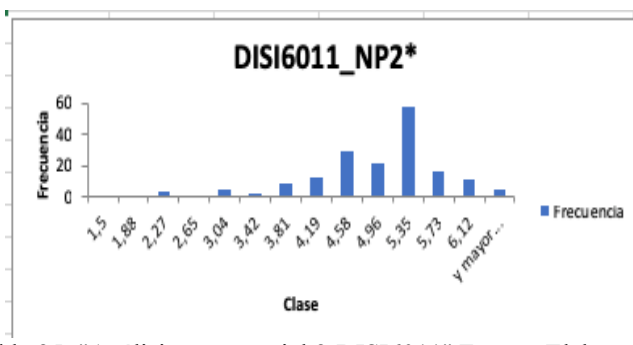


DISI6011_NP1*	
Media	4,71847826
Error típico	0,05361631
Mediana	5
Moda	5
Desviación estándar	0,72728695
Varianza de la muestra	0,52894631
Curtosis	1,88625717
Coefficiente de asimetría	-0,7298878
Rango	4,5
Mínimo	1,5
Máximo	6
Suma	868,2
Cuenta	184

Tabla 24: "Análisis nota parcial 1-DISI6011". Fuente: Elaboración Propia

La Tabla 24, 25 y 26 contiene el análisis descriptivo del ramo "Cultura y Diseño" para las tres primeras notas parciales. Este análisis se hizo con 184 datos para la NP1, 172, datos para la NP2 y 146 datos para la nota parcial 3.

Clase	Frecuencia
1,5	1
1,88	0
2,27	3
2,65	0
3,04	5
3,42	2
3,81	9
4,19	12
4,58	29
4,96	21
5,35	57
5,73	17
6,12	11
y mayor...	11



DISI6011_NP2*	
Media	4,77093023
Error típico	0,06477179
Mediana	5
Moda	5
Desviación estándar	0,84947404
Varianza de la muestra	0,72160615
Curtosis	2,10639948
Coefficiente de asimetría	-1,026128
Rango	5
Mínimo	1,5
Máximo	6,5
Suma	820,6
Cuenta	172

Tabla 25: "Análisis nota parcial 2-DISI6011" Fuente: Elaboración Propia

Tal como se observa en la Tabla 23, existen: 208 datos faltantes para la NP1, 208 para la nota parcial 2 y 209 valores faltantes para la NP3.

El objetivo de conocer estadísticamente el comportamiento de estas tres variables es para utilizar: la media, la desviación estándar y el error, como datos de entrada para el método de imputación de datos a realizar. El método de imputación de datos consistirá en utilizar estos valores, con el objetivo de generar números aleatorios, que respeten la distribución de la muestra.

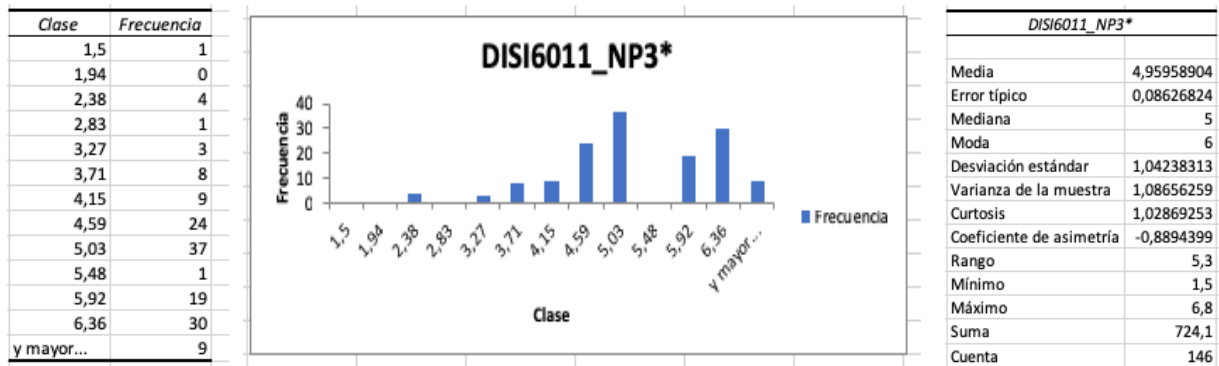


Tabla 26: "Análisis nota parcial 3-DISI6011" Fuente: Elaboración Propia

### DISI6012: Forma y Estructura

La Tabla 27, 28 y 29 contiene el análisis descriptivo del ramo "Forma y Estructura" para las tres primeras notas parciales. Este análisis se hizo con 175 datos para la NP1, 173 datos para la NP2 y 179 datos para la nota parcial 3.

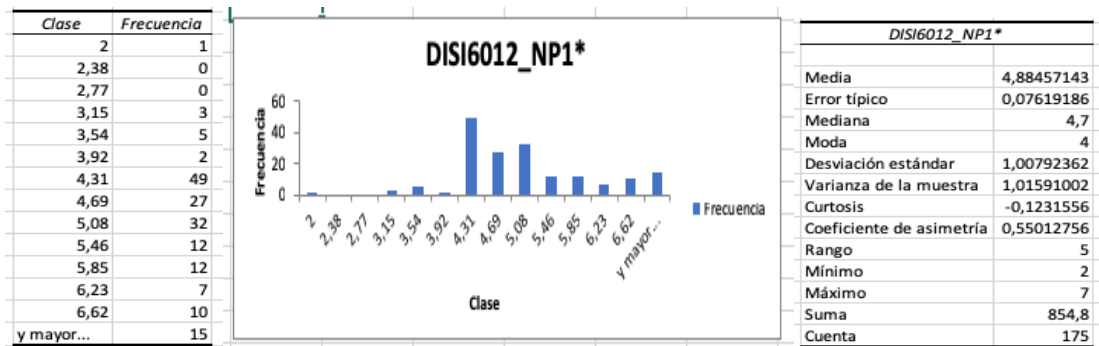


Tabla 27: "Análisis nota parcial 1 - DISI6012". Fuente: Elaboración Propia

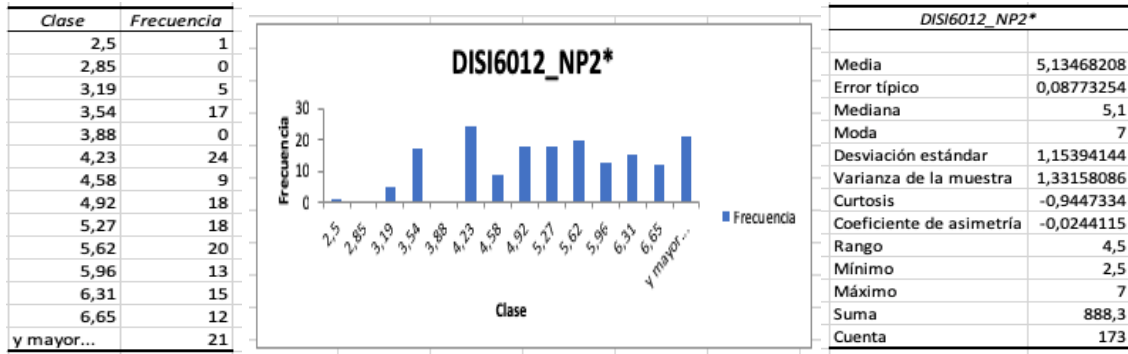


Tabla 28: "Análisis nota parcial 2 - DISI6012". Fuente: Elaboración Propia

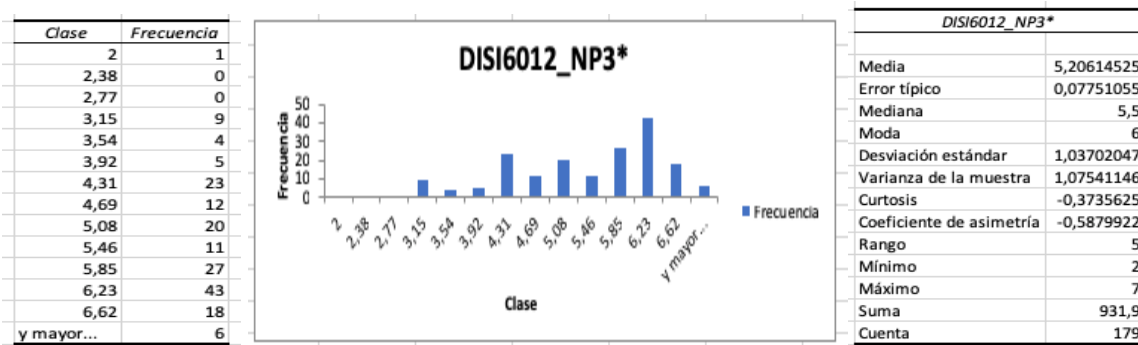


Tabla 29: "Análisis nota parcial 3 - DISI6012". Fuente: Elaboración Propia

Del análisis de la Tabla 23 es posible visualizar que para la NP1 faltan 212 datos y que para la NP2 y NP3 hay 216 datos perdidos.

### DISI6013: Representación Técnica Digital

La Tabla 30, 31 y 32 contiene el análisis descriptivo del ramo “Forma y Estructura” para las tres primeras notas parciales. Este análisis se hizo con 187 datos para la NP1, 170 datos para la NP2 y 37 datos para la nota parcial 3.

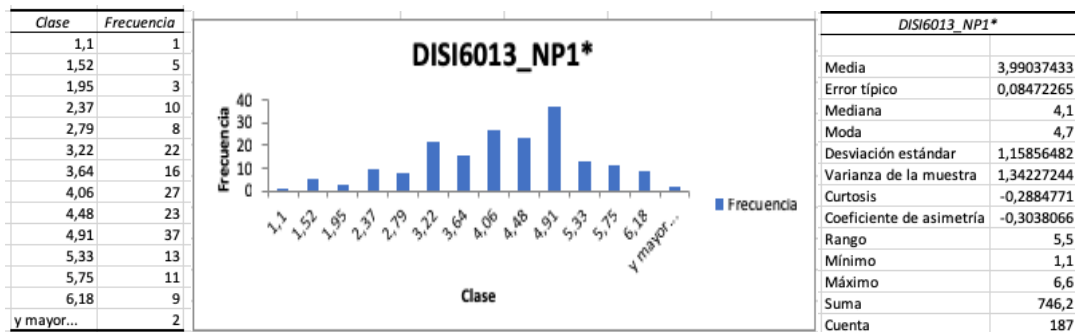


Tabla 30: "Análisis de nota parcial 1 - DISI6013". Fuente: Elaboración Propia

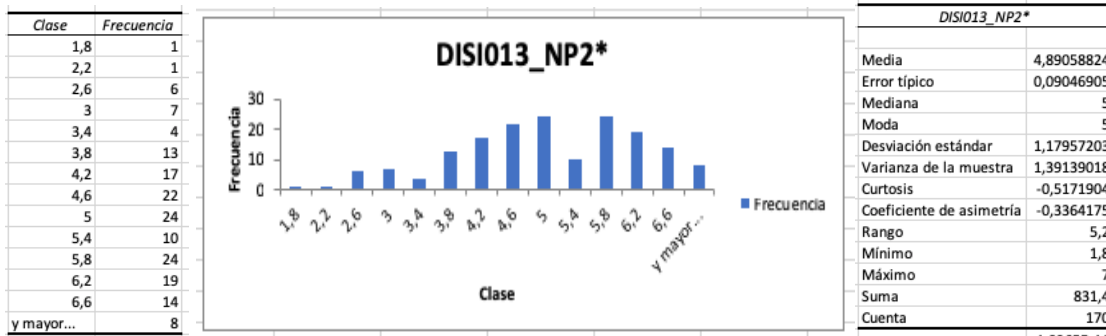


Tabla 31: "Análisis nota parcial 2 - DISI6013". Fuente: Elaboración Propia

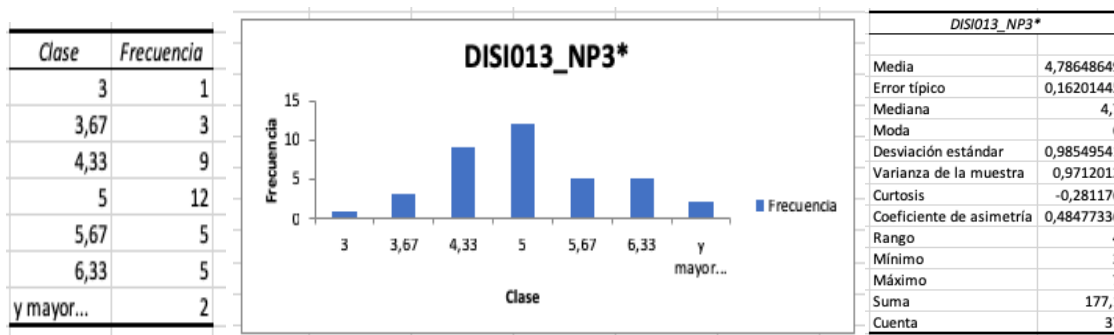


Tabla 32: "Análisis nota parcial 3 - DISI6013". Fuente: Elaboración Propia

Dado que la cantidad de faltos faltantes para la NP3 es demasiado alta, llegando al 88,65% de datos faltantes (ver Tabla 23), es que se toma la decisión de eliminar esta nota del análisis.

### DISI6014: Dibujo I

La Tabla 33, 34 y 35 contiene el análisis descriptivo del ramo “Dibujo I” para las tres primeras notas parciales. Este análisis se hizo con 178 datos para la NP1, 164 datos para la NP2 y 161 datos para la nota parcial 3.

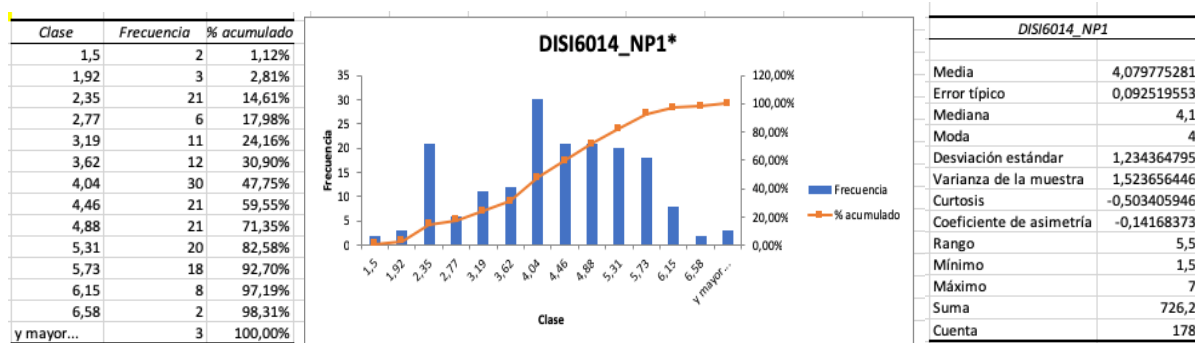


Tabla 33: "Análisis nota parcial 1- DISI6014". Fuente: "Elaboración Propia"



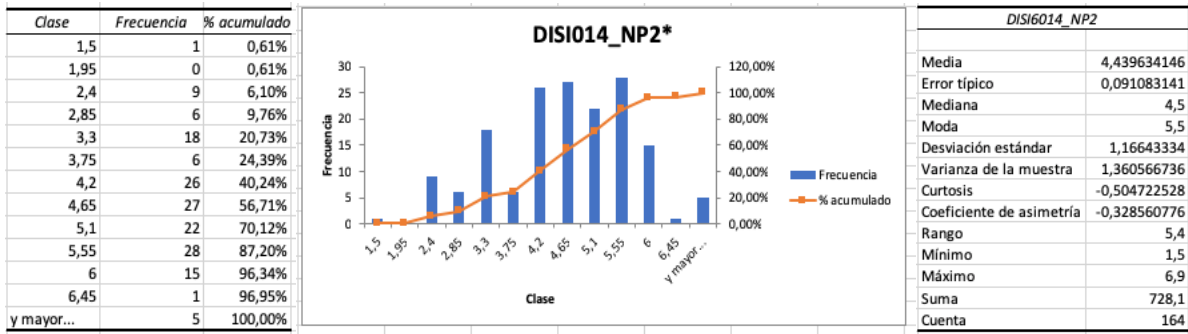


Tabla 34: "Análisis nota parcial 2 - DISI6014". Fuente: Elaboración Propia

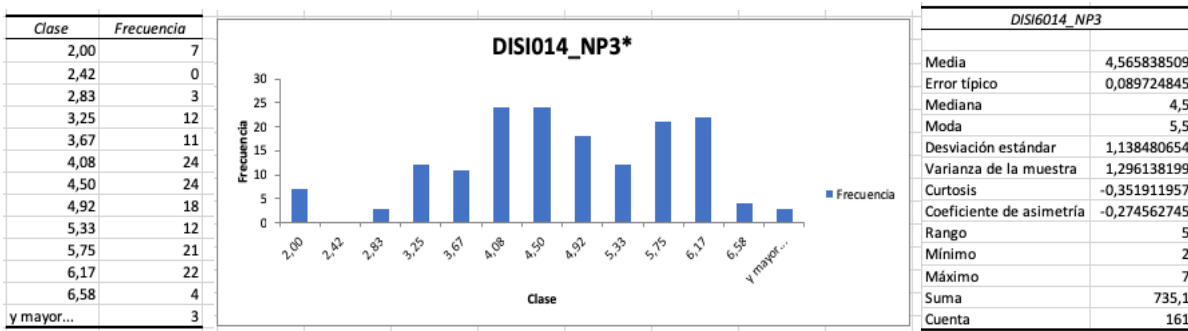


Tabla 35: "Análisis nota parcial 3 - DISI6014". Fuente: Elaboración Propia

Del análisis de la Tabla 23 es posible visualizar que para la NP1 faltan 212 datos y que para la NP2 y NP3 hay 216 datos perdidos.

### DISI6015: Taller de Diseño Industrial I

La Tabla 36, 37 y 38 contiene el análisis descriptivo del ramo "Taller de Diseño Industrial I" para las tres primeras notas parciales. Este análisis se hizo con 203 datos para la NP1, 203 datos para la NP2 y 203 datos para la nota parcial 3.

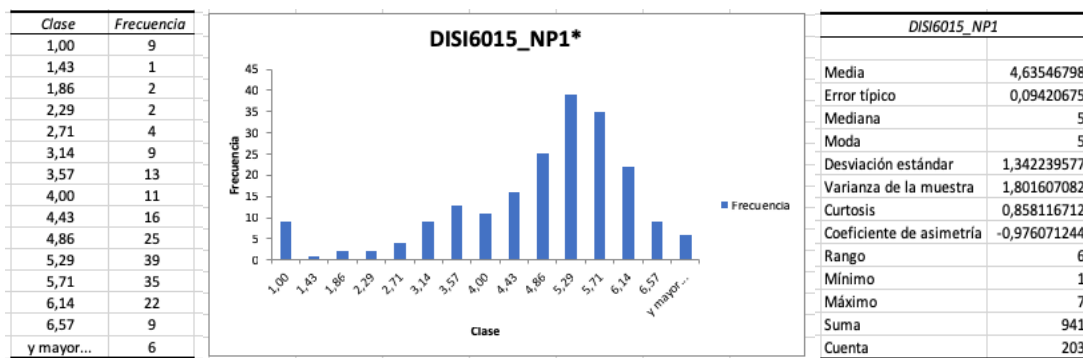


Tabla 36: "Análisis nota parcial 1 - DISI6015". Fuente: Elaboración Propia

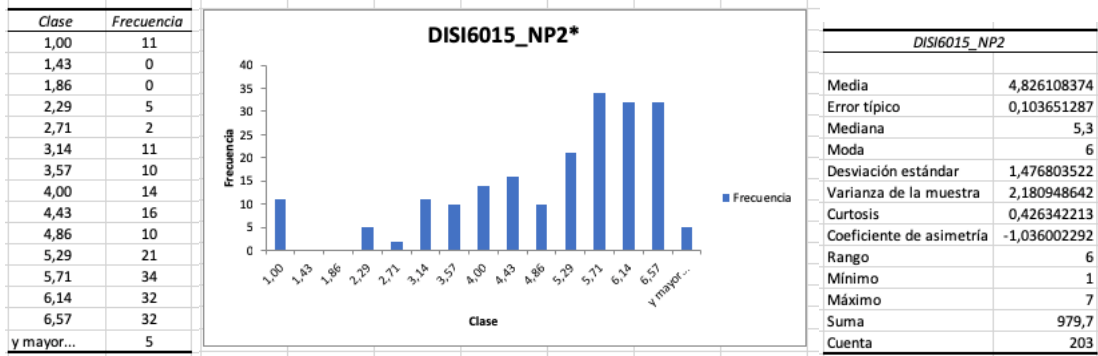


Tabla 37: "Análisis nota parcial 2 - DISI6015". Fuente: Elaboración Propia

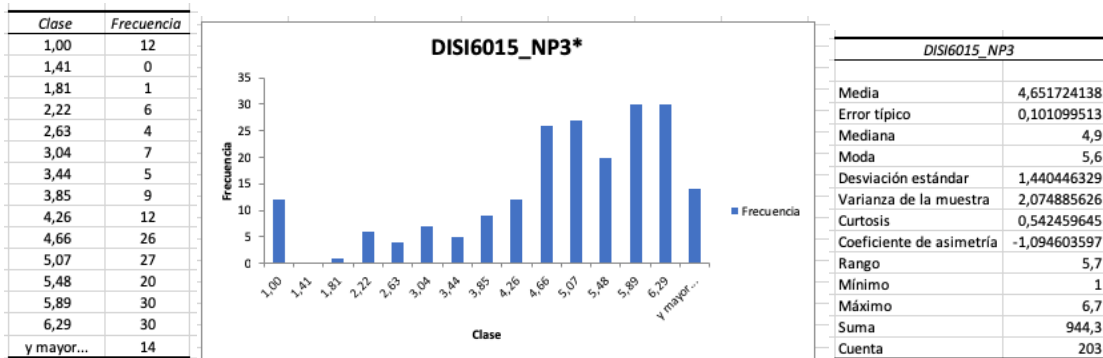


Tabla 38: "Análisis nota parcial 3- DISI6015". Fuente: Elaboración Propia

Del análisis de la Tabla 23 es posible visualizar que para la NP1, la NP2 y la NP3 faltan 211 datos perdidos.

### HUMI6011: Inglés I

La Tabla 39, 40 y 41 contiene el análisis descriptivo del ramo "Inglés I" para las tres primeras notas parciales. Este análisis se hizo con 161 datos para la NP1, 161 datos para la NP2 y 161 datos para la nota parcial 3.

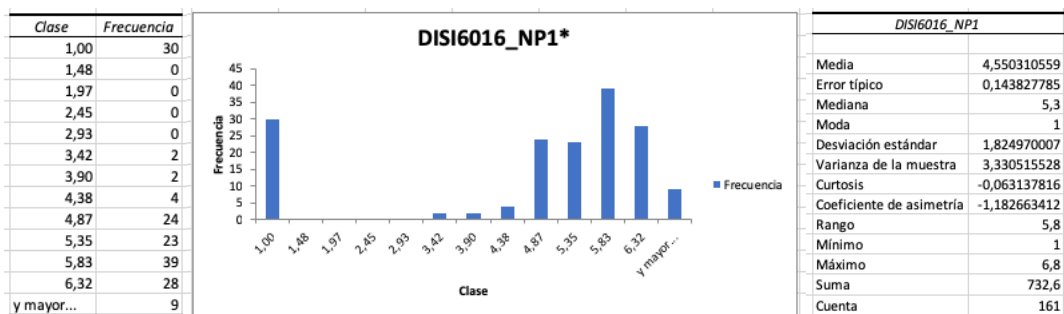


Tabla 39: "Análisis nota parcial 1 – HUMI6011". Fuente: Elaboración Propia

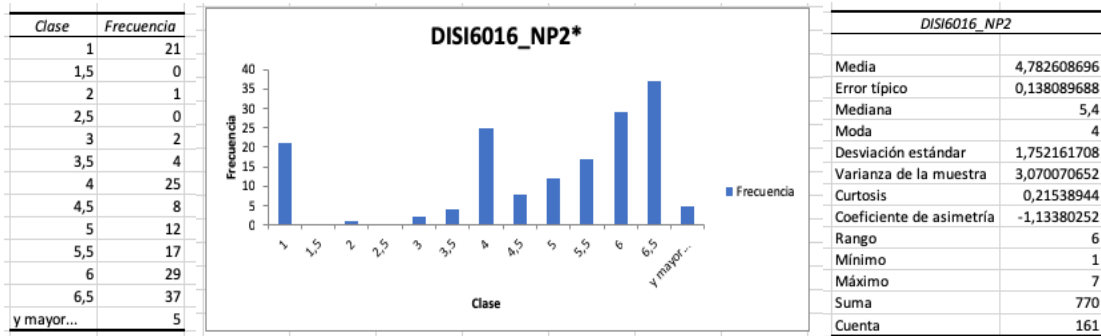


Tabla 40: "Análisis nota parcial 2 – HUMI6011". Fuente: Elaboración Propia

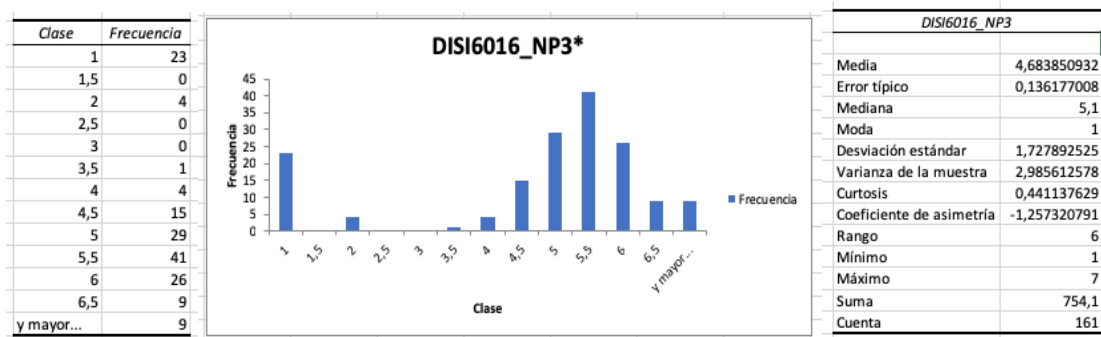


Tabla 41: "Análisis nota parcial 3 - HUMI6011". Fuente: Elaboración Propia

Es posible observar la gran cantidad de notas 1 en las tres evaluaciones parciales. De lo anterior, no es posible identificar si se debe a que los alumnos no contestaron nada correcto o no rindieron la prueba.

### PPSB0001: Taller de Comunicación Efectiva

La Tabla 42 y 43 contiene el análisis descriptivo del ramo “Taller de Comunicación Efectiva” para las dos primeras notas parciales. Este análisis se hizo con 201 datos para la NP1, 199 datos para la NP2.

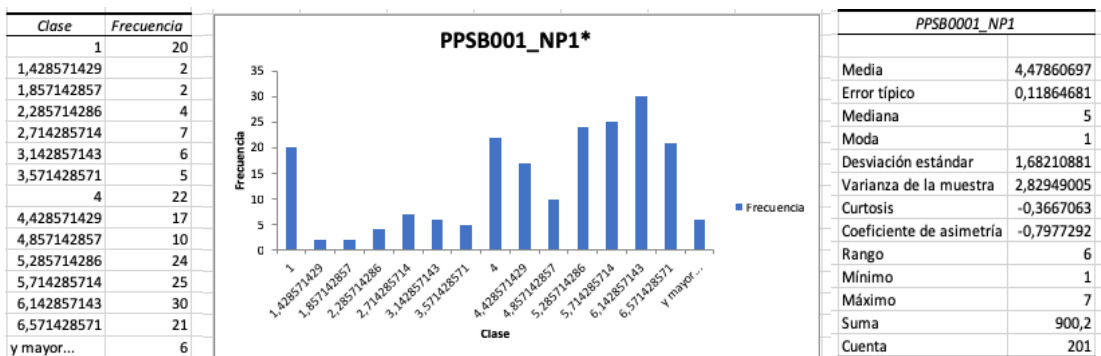


Tabla 42: "Análisis nota parcial 1 - PPSB001". Fuente: Elaboración Propia

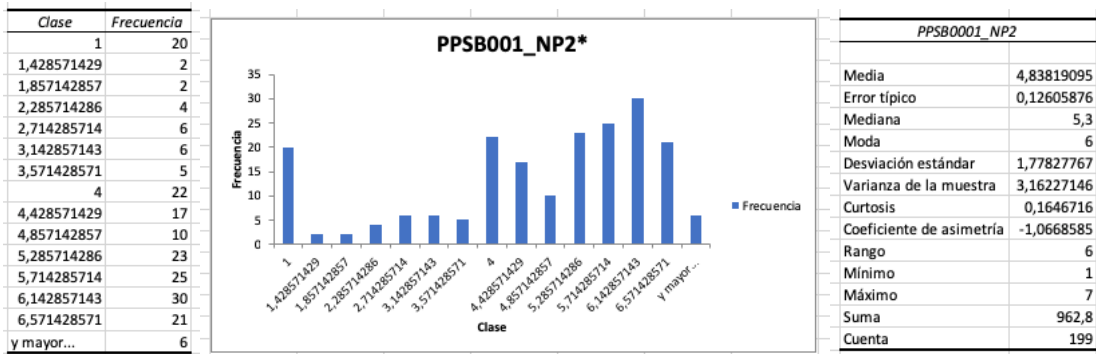


Tabla 43: "Análisis nota parcial 2- PPSB001". Fuente: Elaboración Propia

La nota parcial 3 no fue considerada ya que tal como se muestra en la tabla 9, falta el 96,62% de los datos.

### 5.4.3 Preparación de los datos

Según lo observado en la Tabla 9 y en la distribución de las notas presentadas anteriormente, se pueden observar dos cosas:

- Existe gran cantidad de ausencia de los registros de los datos, lo anterior, se debe a qué no están registrados en los sistemas corporativos.
- Existe gran cantidad de notas 1, lo cual no permite identificar si el alumno no dio la prueba o no respondió nada correctamente.

El conjunto de variables y registros seleccionados debe estar libre de ruidos para no generar sesgo durante su procesamiento. Dado lo anterior, se hace necesario un preprocesamiento de la base de datos para cumplir con un mínimo de calidad de los datos. Si bien, no existe un estándar único de calidad de datos, en este trabajo se considerarán los siguientes:

- **Compleitud:** Inexistencia de datos faltantes (missing values) en los atributos.
- **Consistencia:** El formato y codificación de los valores de un atributo en particular, deben ser idénticos para todos los registros.
- **Coherencia:** Los datos deben responder a reglas lógicas básicas según el contexto de la base. En este punto se deben evaluar los datos identificados como outlier.
- **Validez:** Los registros deben ser coherentes con la organización y/o actividad que se documenta.

Con el objetivo que se cumplan los criterios mencionados anteriormente, el método de imputación de valores perdidos que se utilizará, considerará la distribución de la muestra, generando números aleatorios que tengan la misma media y varianza de la muestra. De esta manera, nos aseguramos que la incorporación de los números que faltan, tiene la misma distribución que la muestra original.

### 5.4.4 Modelamiento

Un modelo es una representación abstracta de los datos y su relación dentro del data set entregado. En este caso, nos interesa encontrar como se relaciona la deserción con las calificaciones obtenidas

en las tres primeras evaluaciones. De esta manera, las variables independientes serán las notas y la variable dependiente (que es la que queremos predecir) si deserta o no deserta.

#### **5.4.4.1 Selección de las técnicas de modelamiento**

RapidMiner cuenta con una gran librería de algoritmos de máquinas de aprendizaje. Para una misma máquina existen distintos algoritmos, cuya configuración está descrita en la sección de ayuda de la plataforma y pueden ser usadas según la necesidad del usuario.

Todos los algoritmos reciben como entrada la base de datos con la que aprenderán y validarán el modelo generado. Sin embargo, para que tengan un correcto funcionamiento, deben ser configurados a través del ingreso de parámetros que varían según el algoritmo de la máquina de aprendizaje.

Se realizarán pruebas con los modelos: Árboles de decisión, Regresión Logística, Redes Neuronales, y Máquinas de Soporte Vectorial.

Dado que este corresponde a un modelo supervisado, es decir, la variable que queremos predecir la conocemos al momento de realizar el estudio, RapidMiner leerá el archivo Excel que contiene el 75% de la información de los estudiantes de primer semestre y con esta información se entrenará el modelo. El 25% restante, se utilizará para validar que tan bien o mal predijo el modelo. Este archivo se utilizará para realizar la predicción.

Con el objetivo de balancear la muestra, se evaluó la cantidad de alumnos que desertaban y no desertaban. Dado que los segundos es un número más reducido, se aplicó la técnica de balanceo ROS con el objetivo de igualar la clase minoritaria. El resultado fue una nueva base de datos, más grande.

#### **5.4.4.2 Generar Diseño de Prueba**

Para realizar el entrenamiento del modelo, se utilizó el set de datos con las calificaciones obtenidas desde el 2010 al 2016, del total de registros, solo se utilizó el 75%. Esto se denomina como primera iteración.

La segunda iteración, consiste en tomar el 25% de los datos restantes, considerados como set de pruebas donde una vez entrenado el modelo, se prueba con los datos 2017.

Para los modelos supervisados el desempeño de éstos se mide a través de la matriz de confusión, más los respectivos indicadores: Recall, Precision, Specificity, Accuracy, Fscore, entre otros.

El método de selección de los modelos, se basa principalmente en el balance que existe entre los indicadores Precision y Recall, donde el indicador Fscore es el promedio entre ambos, dado lo anterior, un buen modelo sería aquel que tenga el mejor Fscore.

#### **5.4.4.3 Construcción de Modelos**

Para la construcción de los modelos señalados se utilizará el software RapidMiner, el cual tiene un grupo bastante amplio de operadores que permiten realizar las técnicas de clasificación para los métodos supervisados.

### 5.4.4.3.1 Árboles de decisión

Los parámetros para un árbol varían según el algoritmo utilizado. Los más utilizados son los siguientes: (1) Máxima Profundidad (Tamaño Poda), (2) Máximo tamaño de separación, (3) Aplicación Poda y (4) Aplicación Prepoda. En este trabajo el parámetro utilizado para optimizar el rendimiento de los árboles será: (1) Máxima profundidad (Maximal Depth) ya que es el criterio que mejor equilibraba el error con el rendimiento del árbol y también dada la cantidad de atributos y de registro, no era necesario aplicar más parámetros.

**Optimal Parameters**  
Maximal Depth: 7

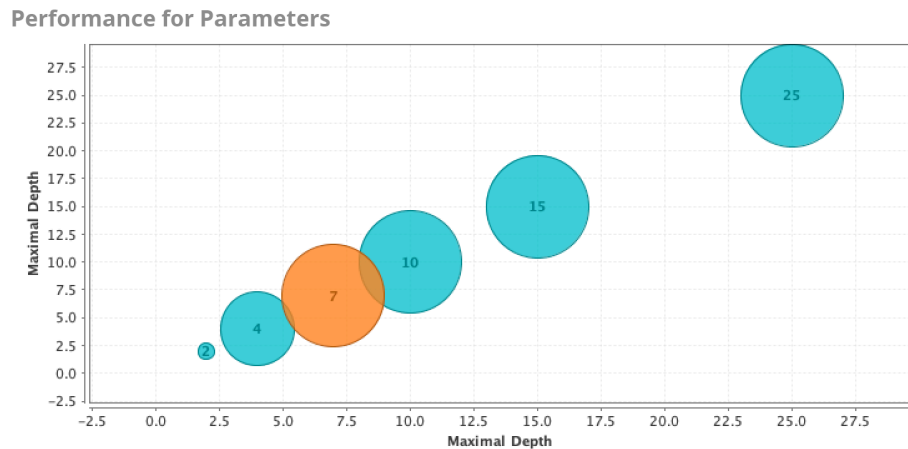


Gráfico 1: Desempeño de los parámetros - Árboles de decisión

Los gráficos de burbujas son una forma de representar datos tridimensionales en forma de un diagrama bidimensional. El Gráfico 1 muestra la relación entre la máxima profundidad, el error asociado al modelo y su desempeño.

Maximal Depth	Performance
2	0,788
4	0,808
7	<b>0,818</b>
10	0,818
15	0,818
25	0,818

Tabla 44: Performance Árbol de Decisión

Con estos parámetros, se obtuvieron los siguientes indicadores:

Indicador	Valor
Accuracy	81,99%
Fscore	70,31%
Recall	59,05%
Precision	89,33%
Specificity	95,00%

Sensitivity	59,05%
Classification Error	18,01%

Tabla 45: Indicadores Árbol de Decisión

La profundidad máxima del árbol, especifica el número máximo de niveles bajo el nodo raíz (el número de veces que se dividirá la muestra repetidamente). Se probó con los siguientes niveles: 2, 4, 7, 10, 15, 25. Con estos valores de máxima profundidad el valor que presentó un mejor desempeño fue 7, con un Accuracy de un 81,99% y un error de 18,01%, tal como se observa en la Tabla 45.

#### 5.4.4.3.2 Random Forest

Para este tipo de modelo de árboles, se utiliza el operador “Random Forest”, donde se busca mejorar los resultados obtenidos por los árboles de decisión. Los parámetros óptimos para este tipo de modelo son: el número de árboles a crear y la profundidad máxima, es decir, el tamaño de la poda.

##### Optimal Parameters

Number Of Trees: **20**  
Maximal Depth: **7**

##### Performance for Parameters

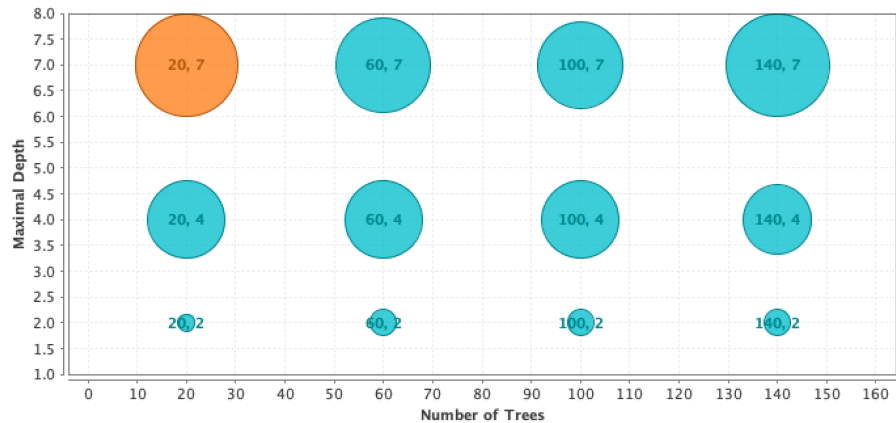


Gráfico 2: Desempeño de los parámetros - Random Forest

El Gráfico 2 indica la relación entre el número de árboles a crear, la máxima profundidad y el error asociado al modelo. Según lo anterior, es posible observar que el mejor desempeño se obtiene con un número de árboles de 20 y una profundidad máxima de 7.

Number of Trees	Maximal Depth	Performance
20	2	0,783
60	2	0,788
100	2	0,788
140	2	0,788
20	4	0,818
60	4	0,818
100	4	0,818
140	4	0,813
<b>20</b>	<b>7</b>	<b>0,833</b>
60	7	0,828

100	7	0,823
140	7	0,833

Tabla 46: Performance Random Forest

Con estos parámetros, se obtuvieron los siguientes indicadores:

<b>Indicador</b>	<b>Valor</b>
Accuracy	82,98%
Fscore	70,30%
Recall	55,71%
Precision	96,00%
Specificity	98,33%
Sensitivity	55,71%
Classification Error	17,02%

Tabla 47: Indicadores Random Forest

Se hicieron 12 pruebas ajustando dos parámetros: la máxima profundidad y la cantidad de árboles a crear. Para la máxima profundidad se utilizaron los siguientes niveles: 2,4 y 7 (los cuales especifican el número máximo de niveles bajo el nodo raíz). Se dejó cada uno de estos parámetros fijos y se fue variando la cantidad de árboles a crear. Se probó con: 20, 60, 100 y 140, tal como se indica en la Tabla 46.

Con estos valores, el modelo que presentó un mejor desempeño fue con un “Número de árboles” igual a 20 y un máximo de profundidad igual a 7, con un Accuracy de un 82,98% y un error de 17,02%.

#### **5.4.4.3.3 Gradient Boosted Trees**

Para este tipo de modelo de árboles, se utiliza el operador “Gradient Boosted Trees”, donde se busca mejorar los resultados obtenidos por los árboles de decisión. Los parámetros óptimos para este tipo de modelo son: el número de árboles a crear y la profundidad máxima, es decir, el tamaño de la poda.



### Optimal Parameters

Number Of Trees: **60**

Maximal Depth: **2**

### Performance for Parameters

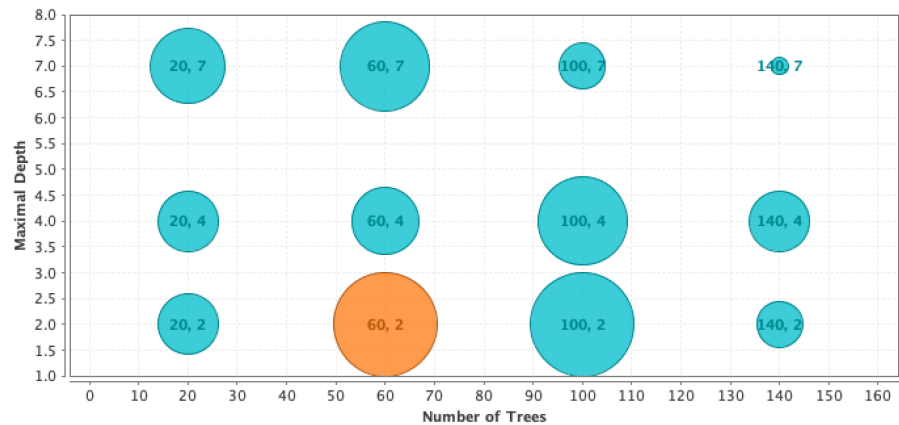


Gráfico 3: Desempeño de los parámetros - Gradient Boosted Trees

Number of Trees	Maximal Depth	Performance
20	2	0,793
<b>60</b>	<b>2</b>	<b>0,823</b>
100	2	0,823
140	2	0,783
20	4	0,793
60	4	0,798
100	4	0,813
140	4	0,793
20	7	0,803
60	7	0,813
100	7	0,783
140	7	0,763

Tabla 48: Performance Gradient Boosted Trees

Con estos parámetros, se obtuvieron los siguientes indicadores:

Indicador	Valor
Accuracy	81,99%
Fscore	72,12%
Recall	64,76%
Precision	82,48%
Specificity	91,67%
Sensitivity	64,76%
Classification Error	18,01%

Tabla 49: Indicadores Gradient Boosted Trees

Se hicieron 12 pruebas ajustando dos parámetros: la máxima profundidad y la cantidad de árboles a crear. Para la máxima profundidad se utilizaron los siguientes niveles: 2,4 y 7 (los cuales especifican el número máximo de niveles bajo el nodo raíz). Se dejó cada uno de estos parámetros fijos y se fue variando la cantidad de árboles a crear. Se probó con: 20, 60, 100 y 140, tal como se indica en la Tabla 48 y en el Gráfico 3.

Con estos valores, el modelo que presentó un mejor desempeño fue con un “Número de árboles” igual a 60 y un máximo de profundidad igual a 2, con un Accuracy de un 81,99% y un error de 18,01%.

#### 5.4.4.3.4 Support Vector Machine

Los parámetros del Support Vector Machine corresponden a los costos asociados al error de clasificación. Por lo tanto, en esta tesis se optimizará el “C” que otorgue el mejor desempeño de los modelos.

Para el ajuste del modelo SVM se necesita determinar dos parámetros, el costo y el parámetro sigma o gamma, asociado a un kernel de tipo Radial Basis Function (RFB). Si bien otros tipos de kernel fueron testeados, acá solo se presentan los resultados obtenidos por uno de tipo RBF que fué el que presentó mejores resultados.

### Kernel Model

Total number of Support Vectors: 95  
Bias (offset): -9.871

w[DISI6013\_NP1] = 10729.026  
w[[DISI6013\_NP2]-[DISI6011\_NP1]\*[DISI6011\_NP1]] = -34702.089  
w[[DISI6011\_NP1]\*[DISI6011\_NP1]] = 45712.921

number of classes: 2  
number of support vectors for class 1: 50  
number of support vectors for class 0: 45

Optimal Parameters  
Gamma: 0.0010000000000000002  
C: 100.0

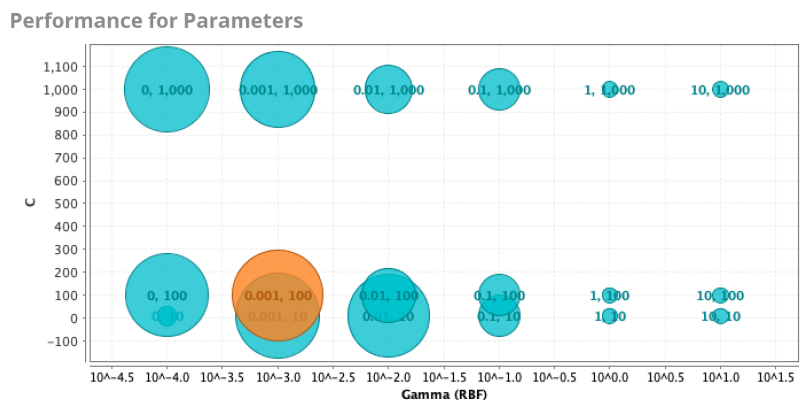


Gráfico 4: Desempeño de los parámetros - SVM

Gamma (RBF)	C	Performance
0,000	10	0,646
0,001	10	0,798

0,010	10	0,793
0,100	10	0,697
1,000	10	0,636
10	10	0,636
0,000	100	0,793
<b>0,001</b>	<b>100</b>	<b>0,813</b>
0,010	100	0,727
0,100	100	0,697
1,000	100	0,636
10	100	0,636
0,000	1000	0,798
0,001	1000	0,778
0,010	1000	0,712
0,100	1000	0,697
1,000	1000	0,636
10	1000	0,636

Tabla 50: Performance Support Vector Machine

Con estos parámetros, se obtuvieron los siguientes indicadores:

<b>Indicador</b>	<b>Valor</b>
Accuracy	81,93%
Fscore	67,76%
Recall	52,86%
Precision	96,00%
Specificity	98,33%
Sensitivity	52,86%
Classification Error	18,07%

Tabla 51: Indicadores Support Vector Machine

El costo C y el gamma son los dos parámetros con los que contamos en los SVM. El parámetro C es el peso que le damos a cada observación a la hora de clasificar un mayor costo, lo cual implicaría un mayor peso de una observación y el SVM sería más estricto

El parámetro gamma establece qué tan rápido caen las similitudes. Si gamma crece implica que la similitud decae más rápidamente. Entonces conforma aumenta la gamma aumenta la varianza, pero el sesgo potencialmente decrece.

Se hicieron 18 pruebas ajustando los parámetros: Gamma y C. Se utilizaron los siguientes parámetros para Gamma: 0,000, 0,001, 0,010, 0,100, 1,000 y 10 y los siguientes parámetros para C: 10, 100 y 1000, tal como se indica en la Tabla 50 y el Gráfico 3.

Con estos valores, el modelo que presentó un mejor desempeño fue con un Gamma de 0,001 y un C de 100, con un Accuracy de un 81,93% y un error de 18,07%.

#### 5.4.4.3.5 Deep Learning

Los algoritmos de redes neuronales varían en la configuración de los parámetros para su correcto funcionamiento. En el caso de este proyecto, se optimizarán dos parámetros principalmente: (1) Tamaño Capa Oculta y (2) Ciclos de Entrenamiento.

El Tamaño de la Capa Oculta indica la cantidad de nodos utilizados en esta capa de la red neuronal. Actualmente existen algoritmos que calculan de manera automática el tamaño ideal de la capa, siendo la mayoría de las veces  $n-1$  nodos, donde  $n$  es la cantidad de variables dependientes. Sin embargo, no siempre este número entrega los mejores desempeños de las redes, por lo que en este trabajo la asignación automática será comparada con asignaciones manuales del número de nodos.

Los ciclos de entrenamiento hacen referencia al número de iteraciones que el algoritmo realizará para ajustar los valores de los pesos  $w_i$  y el coeficiente de distorsión. Mientras mayor número de ciclos asignados, mayor probabilidad de obtener un mejor desempeño de las redes.

Para realizar el modelo de entrenamiento y validación, se utilizaron 4 capas y 15 ciclos de entrenamiento.

Con estos parámetros, se obtuvieron los siguientes indicadores:

Indicador	Valor
Accuracy	78,65%
Fscore	67,25%
Recall	61,43%
Precision	74,95%
Specificity	88,33%
Sensitivity	61,43%
Classification Error	21,35%

Tabla 52: Indicadores Deep Learning

#### 5.4.4.3.6 Logistic Regression

En la etapa de entrenamiento, los algoritmos de regresiones logísticas obtienen las predicciones para un conjunto de observaciones  $X$  y al igual que los support vector machines, es posible calcular el costo de error de las clasificaciones erróneas. Adicionalmente, los algoritmos iteran con el objetivo de identificar el mejor conjunto de los valores  $w_i$ , en donde el costo total de clasificaciones erróneas busca ser minimizado. En este proyecto, el parámetro de las regresiones logísticas será el costo  $C$  para cada error de clasificación

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
DISI6011_NP1	-0.178	-0.286	0.138	-1.290	0.197
DISI6011_NP2	0.203	0.372	0.159	1.276	0.202
DISI6011_NP3	-0.115	-0.222	0.133	-0.862	0.389
DISI6012_NP1	0.165	0.263	0.139	1.190	0.234
DISI6012_NP3	-0.221	-0.372	0.120	-1.833	0.067
DISI6013_NP1	-0.230	-0.330	0.175	-1.315	0.189
DISI6013_NP2	-0.438	-0.861	0.125	-3.500	0.000
DISI6014_NP1	0.126	0.155	0.189	0.665	0.506
DISI6014_NP2	-0.127	-0.203	0.186	-0.678	0.497
DISI6014_NP3	0.082	0.129	0.180	0.457	0.647
DISI6015_NP1	-0.251	-0.274	0.209	-1.200	0.230
DISI6015_NP3	0.051	0.075	0.148	0.342	0.733
HUMI6011_NP1	-0.073	-0.110	0.145	-0.504	0.615
HUMI6011_NP2	0.099	0.164	0.132	0.750	0.453
PPSB0001_NP1	0.097	0.141	0.159	0.612	0.541
PPSB0001_NP2	0.049	0.087	0.126	0.387	0.698
Intercept	2.024	-0.754	-0.491	-4.123	0.000

Tabla 53: Valores Regresión Logística

Con estos parámetros, se obtuvieron los siguientes indicadores:

Indicador	Valor
Accuracy	81,93%
Fscore	68,60%
Recall	55,71%
Precision	92,67%
Specificity	96,67%
Sensitivity	55,71%
Classification Error	18,07%

Tabla 54: Indicadores Logistic Regression

Tal como se observa en la Tabla 39, con los parámetros utilizados, se obtuvo un Accuracy de 81,93% y un Error de Clasificación del 18,07%.

## 5.4.5 Evaluación

### 5.4.5.1 Comparación entre modelos

Accuracy

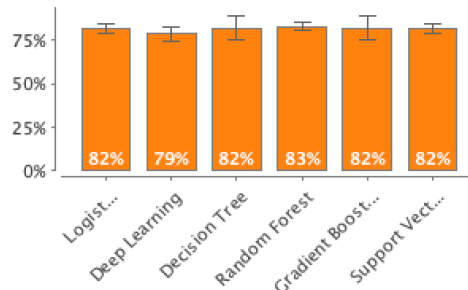


Gráfico 5: Comparación de Accuracy entre Modelos

Classification Error

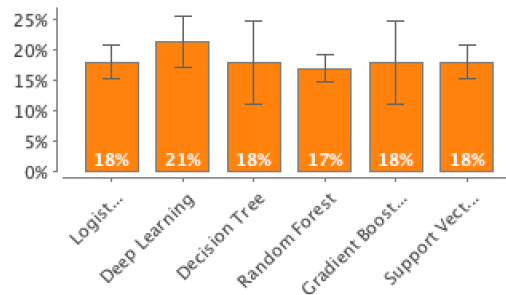


Gráfico 6: Comparación de Errores entre Modelos

Tal como se indica en el Gráfico 5 y 6 es posible identificar que el algoritmo que mejor predice es el Random Forest, logrando una exactitud de un 83% y un error del 17%, lo cual es bastante bajo.

## 5.4.6 Despliegue

La implementación del análisis realizado se basa en la creación de programas de seguimiento a alumnos con el apoyo de modelos matemáticos, que vayan monitoreando si rindió o no la prueba y la nota que obtuvo, debido a que en la actualidad no existe claridad respecto a cuales son las variables que determinan un sistema que indique cuales son los alumnos con mayor probabilidad de hacer abandono de la carrera. Esto permitirá utilizar de manera eficiente y en forma eficaz los recursos.

## CAPÍTULO 6: PROPUESTA DE APOYO TECNOLÓGICO

Esta sección incluye los requerimientos que inciden directamente en el uso del componente de apoyo tecnológico.

### 6.1 ESPECIFICACIÓN DE LOS REQUERIMIENTOS

#### 6.1.1 Requerimientos Funcionales

Los requerimientos funcionales son los que describen cualquier actividad que el sistema deba realizar, en otras palabras, el comportamiento o función particular del sistema cuando se cumplen ciertas condiciones. Para este proyecto en particular, los requerimientos funcionales son los siguientes:

- Permitir el acceso y conexión a las bases de datos institucionales
- Extraer las notas del semestre de interés
- Permitir el procesamiento de información para elaborar modelo
- Permitir el uso de modelos analíticos para la predicción de la deserción.
- Permitir la exportación de los datos procesados con el respectivo scoring del modelamiento para realizar análisis posterior.

#### 6.1.2 Requerimientos No Funcionales

Los requerimientos no funcionales son los que especifican cómo el sistema llevará a cabo la operación y cuales son las especificaciones técnicas con las cuales debe cumplir. Para realizar lo anterior utilizaremos un modelo desarrollado en el año 1987 por Hewlett Packard, en el cual se analizan los factores de calidad de software, bajo el acrónimo de FURPS: funcionalidad (Functionality), usabilidad (Usability), desempeño (Performance) y capacidad de soporte (Supportability). A continuación, se detallan los requerimientos no funcionales para este proyecto de tesis.

Functionality	Requerimientos funcionales indicados previamente
Usability	No aplica
Reliability	Debe tener un sistema que le permita recuperarse ante caídas en un tiempo prudente
Performance	El desempeño debe ser alto, es decir, tener una alta capacidad de procesamiento, de acuerdo a la cantidad de datos que se deben procesar y los modelos analíticos que se utilizarán
Support	Debe tener un soporte similar al resto de los sistemas corporativos de la Universidad.

## 6.2 ARQUITECTURA TECNOLÓGICA

Los sistemas de registro académico de la Universidad han sido desarrollados internamente por el área de Informática. El sistema en funcionamiento en la actualidad, corresponde a una aplicación web (cuyos usuarios principales son alumnos y académicos) desarrollada en PHP (y Javascript) con la base de datos en Oracle. La aplicación fue construida siguiendo un patrón de diseño MVC (que separa lógica de negocios y acceso a datos de interacción con los usuarios en piezas de código distintas), y es desplegada en una arquitectura cliente-servidor.

La plataforma física (servidores y otros equipos) es también administrada por el área de Informática de la Universidad, y conviven en ella distintas plataformas tecnológicas (Windows Server, Linux CentOS), ya sean físicas o virtuales.

## 6.2 DISEÑO DEL SISTEMA VIRTUAL

### 6.2.1 Casos de Uso

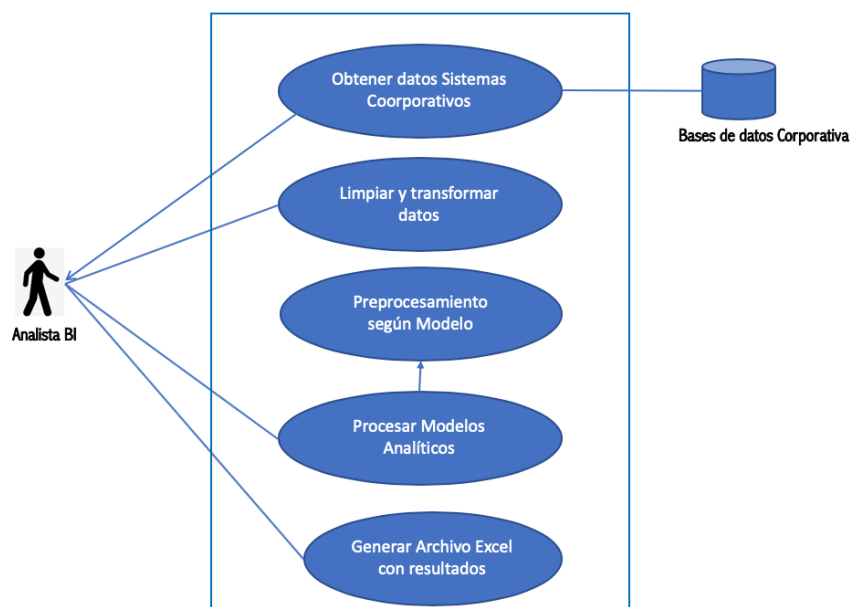


Ilustración 28: Diagrama de Caso de Uso

Caso de uso – Analista BI	Obtener datos Sistemas Corporativos
Descripción	Extraer datos de los sistemas internos
Escenario de Éxito	<ul style="list-style-type: none"> <li>▪ Abrir navegador web</li> <li>▪ Conectarse a Oracle</li> <li>▪ Aplicar queries de extracción de datos</li> <li>▪ Importar datos a estructura CSV</li> </ul>

Caso de uso – Analista BI	Limpiar y Transformar Datos
---------------------------	-----------------------------



Descripción	Preprocesar los datos obtenidos de los sistemas internos en términos de limpieza y transformación
Escenario de Éxito	▪ Cargar datos a Excel
	▪ Generar un XLS transitorio
	▪ Aplicar criterios para imputar y limpiar datos
	▪ Generar archivo XLS con limpieza realizada

Caso de uso – Analista BI	Procesar Modelos Analíticos
Descripción	Aplicar modelos analíticos seleccionados al DataSet limpio y preprocesado según el tipo de modelo
Escenario de Éxito	▪ Incluye Preprocesamiento según Modelo
	▪ Se carga el Data Set preprocesado según modelo
	▪ Se separa una cantidad para entrenamiento y otra para testear
	▪ Se utiliza el set de datos de entrenamiento para el aprendizaje del modelo
	▪ Se aplica modelo entrenado al set de testeo
	▪ Se obtienen los resultados

Caso de uso – Analista BI	Preprocesar Modelo Analíticos
Descripción	Preparar Data Set en base al modelo que se quiere aplicar
Escenario de Éxito	▪ Se carga el set de datos limpios y transformados en RapidMiner
	▪ Se especifica el Modelo a utilizar
	▪ Se revisan las necesidades de acuerdo al modelo seleccionado
	▪ Se realizan las depuraciones y transformaciones correspondientes al Modelo
	▪ Se genera nuevo Data Frame preprocesado en el repositorio local de RapidMiner
▪ Generar archivo XLS con resultados	

Caso de uso – Analista BI	Generar archivo Excel con resultados
Descripción	Generar listado con los resultados de los modelos y tomar medidas con aquellos alumnos que tengan probabilidad de desertar
Escenario de Éxito	▪ Evaluar resultados

	<ul style="list-style-type: none"> <li>▪ Generar informes con alumnos con probabilidad de desertar</li> </ul>
	<ul style="list-style-type: none"> <li>▪ Enviar a los Jefes de Carrera</li> </ul>

### 6.2.2 Diagrama de arquitectura del sistema

En la figura que se presenta a continuación se observa el diagrama de arquitectura del sistema, donde se plasma en forma gráfica lo que se quiere construir y obtener.

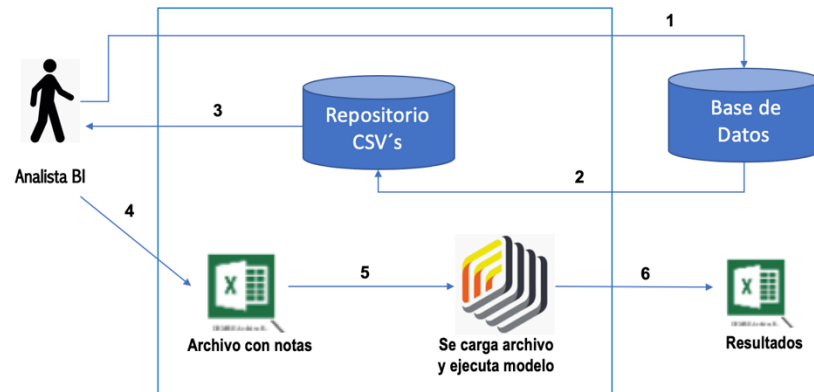


Ilustración 29: "Arquitectura Tecnológica". Fuente: Elaboración Propia

Para este proyecto de tesis, se utilizaron 3 herramientas tecnológicas: Rapid Miner, Excel y los Sistemas de Información Corporativos de la Universidad. Los pasos a seguir se describen a continuación:

1. Conectarse al servidores de “Base de Datos” mediante dirección IP.
2. Elegir la base de datos específica de la Asignatura de la cual se quiere extraer las notas.
3. Mediante un código PL-SQL y herramientas del IDE se accede a los datos y se extraen en el formato requerido, en este caso, en formato Excel.
4. Se carga el archivo Excel en el software RapidMiner.
5. Se ejecuta el modelo de predicción desarrollado.
6. Se extraen los resultados en un archivo Excel para realizar seguimiento a los alumnos identificados.

Finalmente, se realiza un reporte con las estrategias que se van a implementar para nivelar y apoyar a los alumnos con bajo rendimiento y alta probabilidad de desertar.

### 6.2.3 Diagrama de despliegue

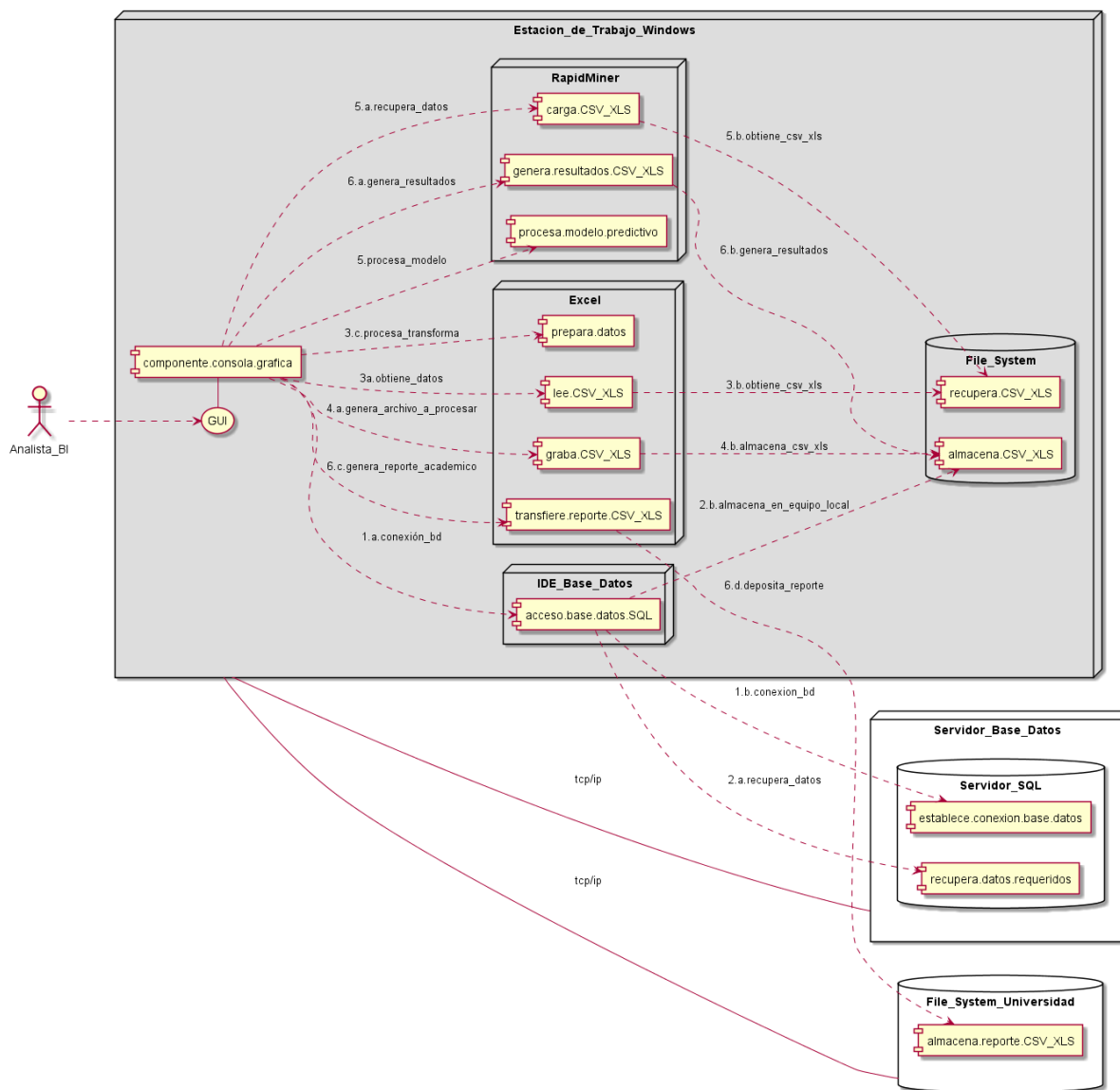


Ilustración 30: Diagrama de despliegue Arquitectura para el Diseño - Elaboración Propia

La Ilustración 30 muestra el diagrama de despliegue con el objetivo de identificar la estructura física de la arquitectura propuesta. El actor “Analista BI” se conecta al servidor de base de datos para extraer la información de las asignaturas que se desea analizar. Luego esta información se almacena en el equipo local y se comienza con el procesamiento de datos. Una vez que los datos están en el formato adecuado, se cargan en el software RapidMiner y se comienza el análisis. Una vez que se generan los resultados, estos son analizados y se genera un reporte académico.

## **CAPÍTULO 7: GESTIÓN DEL CAMBIO**

### **7.1 CONTEXTO DE LA ORGANIZACIÓN**

La ACME es una institución de educación superior del Estado de Chile, sin fines de lucro acreditada en 2016 por cuatro años en las áreas de Gestión Institucional, Docencia de Pregrado y Vinculación con el Medio.

Su misión es generar, desarrollar, integrar y comunicar todas las áreas del conocimiento y cultura. Lo anterior, constituye el fundamento de las actividades de la Universidad, asumiendo la formación de personas y la contribución al desarrollo cultural y material de la Nación.

Cumple su misión a través de las funciones de docencia, investigación y creación en las ciencias y las tecnologías, las humanidades y las artes, y de extensión del conocimiento y la cultura en toda su amplitud, procurando ejercer estas funciones con el más alto nivel de exigencia.

Es responsabilidad de la Universidad contribuir con el desarrollo del patrimonio cultural y la identidad nacionales y con el perfeccionamiento del sistema educacional del país.

### **7.2 ANÁLISIS DE LOS PRINCIPIOS DE DISEÑO**

Los principios analizados en este apartado están basados en el Modelo Integral de Liderazgo y Gestión del Cambio (Olguín, Crawford, & Soto, 2016). Este modelo incorpora el análisis de diferentes dominios a considerar para la realización de un plan de gestión del cambio.

#### **7.2.1 Liderazgo y gestión del proyecto de cambio**

El proyecto de tesis fue liderado por un equipo del área de Estadísticas de la Universidad, ya que desde el departamento de Gestión Académica Institucional se decidió que esta área sea quien maneje los datos académicos de la Universidad y lidere esta iniciativa.

#### **7.2.2 Estrategia y sentido del proceso de cambio**

El sentido del cambio nace al interior de la organización, a partir de la necesidad de conocer las razones por las cuales los alumnos desertan. Lo anterior se enmarca en un proyecto de mejoramiento institucional que tiene como propósito mejorar el desempeño de distintas unidades, en particular, la unidad de Gestión Académica Institucional.

#### **7.2.3 Cambio y Conservación**

El cambio si bien no es complejo de realizar, se visualiza difícil de implementar debido a la resistencia que tienen algunos docentes en ajustarse a los nuevos protocolos que este monitoreo de alumnos requiere. Se entiende y comparte la importancia, pero finalmente es un cambio conductual que afecta directamente las prácticas de trabajo que los involucra en el día a día. En las reuniones se percibe el interés, sin embargo, es la sobrecarga académica lo que se reclama y dificulta el correcto operar. El proceso de enseñanza aprendizaje sigue siendo el mismo, la diferencia está en el plazo que se está imponiendo a los docentes para ingresar y registrar las notas, una vez que se ha realizado la evaluación.

#### **7.2.4 Organización y Estructura del proyecto de cambio**

Los principales actores en el proyecto son: el jefe de Gestión Académica Institucional, Jefe de la Unidad de Estadísticas, Jefe del proyecto de Mejoramiento Institucional.

El flujo de información y conversaciones comienza desde la Unidad de Gestión Académica, donde se realizan análisis permanentes de la gestión de la Institucional y se evalúan las oportunidades de mejora en relación a las prioridades e intereses de la Universidad.

#### **7.2.5 Gestión Emocional**

Las personas que trabajan en la Universidad poseen un fuerte compromiso social y en general, son personas que llevan mucho tiempo trabajando ahí. Conocen el funcionamiento organizacional y han pasado por varias administraciones donde se han implementado varios cambios tecnológicos que han implicado adquirir nuevas capacidades y conocimientos, en particular, a los funcionarios administrativos de la Universidad. Este cambio, es una de las pocas veces que no involucra a los funcionarios sino a los docentes. Por otra parte, ha generado muy buena recepción el hecho de que se generen trabajos para los alumnos que tienen mejores notas y cumplirían el rol de tutores.

#### **7.2.6 Comunicaciones**

El proyecto ha generado innumerables instancias de comunicación y conversación, donde diversos actores han podido dar sus puntos de vista al respecto y ser considerados dentro del desarrollo del sistema de alertas tempranas.

#### **7.2.7 Desarrollo de Habilidades**

Las habilidades a incorporar y gestionar se separan en dos grupos objetivos: los docentes y la Unidad de Estadística. Los docentes, deben desarrollar el hábito de ingresar las notas y controlar la asistencia (actividad que hoy en día no ocurre) y luego dejarlas registradas en un sistema que será el que alimente con información a la Unidad de Estadística para realizar su labor. Por otra parte, la Unidad de Estadística, tendrá una nueva función dentro de sus procesos, el cual consistirá en realizar el seguimiento a los alumnos.

#### **7.2.8 Gestión del Poder**

En la Universidad las fuentes de poder están en las jefaturas a cargo de cada departamento, y a su vez, cada uno de estos cargos, sigue las instrucciones y decisiones que tome el Decano, él cual, a su vez, sigue las decisiones que tome el Rector. Los funcionarios, tienen poco poder de generar cambios, esto radica más en los alumnos los cuales por medio de movilizaciones pueden detener el funcionamiento de la Universidad. En este proyecto, se debe integrar a los decanos y jefes a su cargo, para transmitir la importancia de contar con la información anteriormente mencionada de manera oportuna.

#### **7.2.9 Monitoreo y Evaluación del Proceso**

El monitoreo del proyecto tendrá dos líneas de seguimiento. Una tiene relación con el ingreso de las notas y asistencia a los sistemas y por otra parte la evaluación de los alumnos con probabilidad de desertar. Ambos procesos están íntimamente relacionados, ya que, si no se realiza el ingreso de las

notas de manera oportuna, no se podrá realizar un seguimiento apropiado ya que podría suceder que si la información no se tiene a tiempo, los alumnos ya se hayan retirado.

La evaluación del proceso se medirá en función del aumento de la tasa de retención, lo cual a su vez depende, del ingreso oportuno de la información que se necesita analizar.

#### **7.2.9.10 Inicio, hitos, ritos y cierre**

Un aspecto importante dentro del proyecto es comunicar las etapas y mostrar los beneficios que obtendrá la Universidad con su implementación. Dentro de los hitos relevantes se encuentra mostrar los resultados del estudio para que los académicos y autoridades tomen consciencia de la ausencia de datos y cómo ésta influye en la falta de acciones preventivas.

### **7.3 CARACTERIZACIÓN DEL CAMBIO**

Este proyecto tiene el desafío de incorporar nuevas prácticas de trabajo en los docentes, monitorear el desempeño de los alumnos por parte de la Unidad de Estadística e ir recogiendo nuevas fuentes de información que puedan describir de mejor manera el fenómeno de la deserción. Lo que se visualiza como inmediato, es la difusión de los beneficios que tiene para la Universidad el ingreso oportuno de esta información y de los costos que tiene el no hacerlo.

### **7.4 FACTORES CRÍTICOS DE ÉXITO**

La variable crítica de éxito en este proyecto es el ingreso oportuno de la información, ya que, si esta es ingresada a tiempo, es posible realizar seguimiento a los alumnos con probabilidad de desertar y aplicar la nivelación que necesitan por medio de la intervención de tutores.

### **7.5 PLAN DE GESTIÓN DEL CAMBIO**

Las principales acciones a tomar para gestionar el cambio son las siguientes:

- Reuniones a nivel de decanos y jefes de departamentos para mostrar los resultados del estudio y los beneficios que obtendría la Universidad ingresando a tiempo las notas de las asignaturas.
- Comenzar con aquellas carreras que tienen un índice de deserción mayor y luego ir incorporando las que tienen tasas menores.
- Crear un protocolo que ayude a los jefes de carrera a monitorear el ingreso de las notas parciales.
- Crear un protocolo a los docentes dando a conocer los plazos.
- Crear un plan de incentivo para estimular el ingreso de las notas a tiempo.
- Reuniones semestrales con las autoridades correspondientes para evaluar el plan de acción y analizar las acciones preventivas o correctivas que son factibles de realizar.

## CAPÍTULO 8: EVALUACIÓN DEL PROYECTO

En este capítulo se detalla la evaluación económica del proyecto, así cómo también, se realiza una evaluación del modelo predictivo con resultados de la retención de alumnos del primer semestre del 2018

### 8.1 PLAN PILOTO

Tal como se mencionó en el punto 5.4.2.1 los datos que se utilizaron para entrenar y testear el modelo contenían las notas de primer año desde el 2010 al 2016. Por lo tanto, para probar la validez del modelo, se decidió utilizar los datos del 2017 y 2018.

Dado que el modelo que obtuvo un mejor rendimiento fue el Random Forest en la fase de entrenamiento, se utilizará este mismo algoritmo para probar la eficiencia del modelo.

La base de datos del 2017 y 2018 contenía 125 registros. El tratamiento de datos faltantes fue el mismo utilizado con la base de datos de entrenamiento, es decir, se hizo un análisis estadístico de los datos entregados para cada asignatura y luego, considerando la media y la desviación estándar de los valores promedios, se generaron números aleatorios considerando los estadísticos descriptivos recién mencionados. Una vez hecho esto, se procedió aplicar el algoritmo.

Los resultados obtenidos se describen a continuación:

#### Random Forest – Optimal Parameters

**Optimal Parameters**  
 Number Of Trees: **20**  
 Maximal Depth: **4**

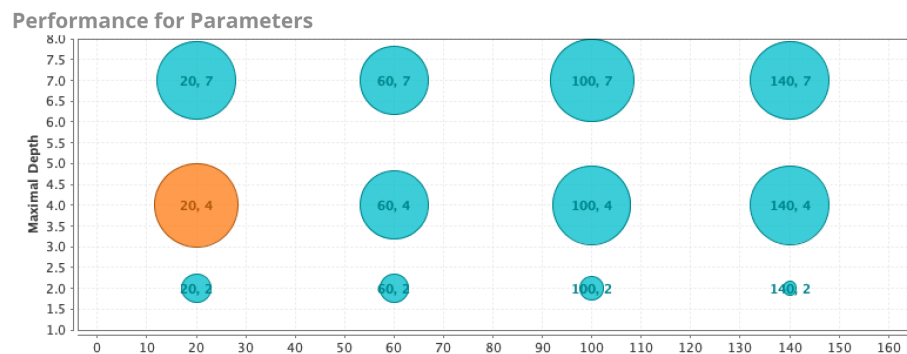


Gráfico 7: Desempeño de los parámetros - Random Forest

Number of Trees	Maximal Depth	Performance
20	2	0,793
60	2	0,793
100	2	0,788
140	2	0,778
<b>20</b>	<b>4</b>	<b>0,848</b>
60	4	0,833
100	4	0,843

140	4	0,843
20	7	0,843
60	7	0,833
100	7	0,848
140	7	0,843

Tabla 55: Performance Random Forest - Prueba Piloto

<b>Indicador</b>	<b>Valor</b>
Accuracy	74,50%
Fscore	61,26%
Recall	57,14%
Precision	69,45%
Specificity	84,70%
Sensitivity	57,14%
Classification Error	25,50%

Tabla 56: Indicadores Random Forest - Prueba Piloto

Se realizaron 12 pruebas ajustando dos parámetros: la máxima profundidad y la cantidad de árboles a crear. Para la máxima profundidad se utilizaron los siguientes niveles: 2,4 y 7 (los cuales especifican el número máximo de niveles bajo el nodo raíz). Para cada uno de estos niveles se fue variando la cantidad de árboles a crear. Se probó con: 20, 60, 100 y 140, tal como se indica en la Tabla 56 y en el Gráfico 7.

Con estos valores, el modelo que presentó un mejor desempeño fue con un “Número de árboles” igual a 20 y un máximo de profundidad igual a 4, con un Accuracy de 74,50% y un error del 25,50%.

Si comparamos los valores del ciclo de entrenamiento con los valores del plan piloto, vemos que estos varían, tal como se muestra en la tabla 47, la cual se muestra a continuación:

<b>Indicador</b>	<b>Valor</b>	<b>Indicador - Piloto</b>	<b>Valor - Piloto</b>
Accuracy	82,98%	Accuracy	74,50%
Fscore	70,30%	Fscore	61,26%
Recall	55,71%	Recall	57,14%
Precision	96,00%	Precision	69,45%
Specificity	98,33%	Specificity	84,70%
Sensitivity	55,71%	Sensitivity	57,14%
Classification Error	17,02%	Classification Error	25,50%

Tabla 57: Comparación valores Random Forest

Al comparar ambos resultados, se puede observar que los valores obtenidos en el plan piloto, son menores en cuanto a la exactitud con la cual el modelo predice y también el error en la clasificación es más alta, sin embargo, lo anterior no significa que los resultados sean malos y que tengan bajo poder de predicción.



Dado que la cantidad de datos no es un volumen tal alto, es posible que el modelo al momento de entrenarse y testear los resultados se haya sobreajustado, por lo tanto, al aumentar el número de registros este sobreajuste cambia ya que tiene una cantidad de datos mayor con la cual trabajar.

La Clase que nos interesa predecir es la Clase 1, es decir, aquellos alumnos que DESERTAN. La Clase 0 es aquella que NO DESERTA.

El objetivo es poder encontrar cuales son los factores predictores y cuales son los valores que estos toman para poder tomar medidas preventivas sobre un grupo de alumnos que presenta un mal desempeño académico. Dado lo anterior el modelo nos entrega la siguiente información.

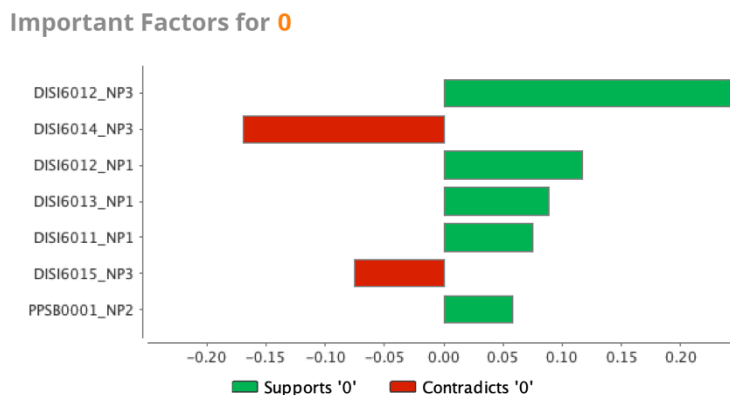


Gráfico 8: Simulación Alumnos que NO DESERTAN

El Gráfico 3 se lee de la siguiente manera: Los factores que definen que un alumno NO deserte son los siguientes:

- DISI6012\_NP3
- DISI6012\_NP1
- DISI6013\_NP1
- DISI6011\_NP1
- PPSB001\_NP2

Por el contrario, los factores que determinan que un alumno deserte son:

- DISI6014\_NP3
- DISI6015\_NP3

Lo anterior tiene sentido, ya que DISI6014 corresponde a la asignatura de “Dibujo I” y DISI6015 corresponde a la asignatura “Taller de Diseño Industrial I” que son los ramos principales de la carrera.

A continuación, se presenta la predicción que realizó el modelo respecto a las notas que podrían a los alumnos en esta condición:

<b>Código</b>	<b>Asignatura</b>	<b>Nota</b>
DISI6012_NP3	Forma y Estructura	4,409
DISI6012_NP1	Forma y Estructura	4,369
DISI6013_NP1	Representación técnica digital	3,193
DISI6011_NP1	Cultura y Diseño	3,664
PPSB001_NP2	Comunicación Efectiva	4,581
DISI6014_NP3	Dibujo I	3,701
DISI6015_NP3	Taller de Diseño Industrial I	4,398

Tabla 58: Simulación Notas Random Forest

La Tabla 59 indica lo siguiente: los alumnos que no desertan son aquellos que tienen notas superiores a las indicadas en la comuna “Nota”, en las asignaturas DISI6011, DISI6012, DISI6013 y PPSB001; por el contrario, los alumnos que tienen notas iguales o inferiores en la nota parcial 3 en las asignaturas DISI6014 Y DISI6015, tienen mayor probabilidad de desertar, ya que estos son los ramos principales de la carrera.

## 8.2 DEFINICIÓN DE BENEFICIOS Y COSTOS

Para calcular los costos asociados que tiene para la Universidad la deserción de alumnos, se va identificar los ingresos que se dejan de percibir por concepto de arancel y se considerará los años que quedan para finalizar la carrera. En este caso, la duración es de 5 años.

Los datos para realizar el cálculo mencionado se presentan a continuación:

<b>Año</b>	<b>2013</b>	<b>2014</b>	<b>2015</b>	<b>2016</b>	<b>2017</b>
% de deserción	25, 38%	27,50%	27,89%	31,31%	33,99%

Tabla 59: "Porcentaje de deserción de alumnos en la carrera de Diseño en Industrial"

En la tabla 60 es posible observar que existe una visible tendencia al alza desde el año 2013, la cual se ha mantenido hasta la fecha. Los valores registrados en cada celda son el resultado de:

- Alumnos que se retiran durante el primer año, es decir, alumnos que ingresan el año “t” y dejan de asistir en algún momento del año académico. Dejan de registrar asistencia a clases. Algunos formalizan su salida, otros simplemente dejan de asistir.
- Alumnos que no se matriculan y/o no inscriben ramos el año siguiente, es decir, ingresan el año “t”, sin embargo, no inscriben ramos y/o no se matriculan el año “t+1”.
- Alumnos que se retiran antes de ser eliminados de la carrera ya que no presentan un buen rendimiento académico.

La Universidad tiene capacidad para recibir 100 alumnos para la carrera de Diseño cada año, por lo tanto, si consideramos el valor del arancel y la cantidad de alumnos que desertan, podemos cuantificar el costo que significa para la Universidad.

<b>Año</b>	<b>2012</b>	<b>2013</b>	<b>2014</b>	<b>2015</b>	<b>2016</b>	<b>2017</b>
Nº de desertores	26	28	28	32	34	38

Tabla 60: "Cantidad de alumnos que desertan por año"

Considerando los datos de la tabla anterior, tenemos que: el año 2013, 26 alumnos desertaron de la carrera; el 2014 fueron 28 alumnos, pero dado que el año anterior fueron 26, los alumnos desertores serían 54. Continuando con el mismo cálculo, para los siguientes años nos da lo siguiente:

<b>Año</b>	<b>2012</b>	<b>2013</b>	<b>2014</b>	<b>2015</b>	<b>2016</b>	<b>2017</b>
Nº de desertores	26	28	28	32	34	38
		26	28	28	32	34
			26	28	28	32
				26	28	28
					26	28
						26
<b>Total</b>	<b>26</b>	<b>54</b>	<b>82</b>	<b>114</b>	<b>148</b>	<b>186</b>

Tabla 61: "Cantidad acumulada de alumnos que desertan"

<b>Arancel</b>	<b>Año de referencia</b>
2012	\$3.521.357
2013	\$3.686.569
2014	\$3.790.600
2015	\$3.938.500
2016	\$4.187.900
2017	\$4.387.400

Tabla 62: "Arancel de referencia por año"

Tomando en consideración los datos anteriores y multiplicando la cantidad de alumnos que se retiran cada año por el arancel que se deja de pagar, nos da los siguientes valores:

<b>Año</b>	<b>Ingreso que se deja de percibir</b>
2012	\$91.555.282
2013	\$194.779.214
2014	\$300.916.014
2015	\$426.948.014
2016	\$569.336.614
2017	\$736.057.814

Tabla 63: "Ingresos que deja de percibir la Universidad por concepto de deserción de alumnos"

Realizando una lectura de la tabla 62, podemos decir que al año 2016, tenemos 5 años acumulados de alumnos que desertan, lo cual corresponde a un total de 148 alumnos. Lo anterior equivale a decir que la Universidad deja de percibir \$569.336.614 por concepto de arancel.

### **Inversión inicial**

Para realizar el cálculo de inversión inicial, solamente se tomará en consideración las horas hombres (HH) de las personas que forman parte del proceso de toma de decisiones y de quienes están encargados de ejecutar aquellas decisiones.

<b>Stakeholders</b>	<b>Remuneración Aproximada</b>	<b>% de dedicación</b>	<b>Remuneración Proyecto</b>
Director departamento Diseño	\$4.500.000	5	\$225.000
Jefe de Gestión Académica Diseño	\$2.400.000	20	\$480.000
Coordinador Académico	\$1.700.000	15	\$255.000
Consultor externo	\$1.500.000	60	\$900.000
		Valor mensual	\$1.860.500
		Total 3 meses	\$5.581.500

### **Costos de mantención de la solución**

La aplicación requiere actualizaciones mensuales que demandan muy pocas horas de trabajo.

<b>Stakeholder</b>	<b>Remuneración aproximada</b>	<b>% de dedicación</b>	<b>Remuneración proyecto</b>
Jefe de la Unidad de Estadística	\$1.500.000	5	\$75.000
		Valor Mensual	\$75.000
		Total 12 meses	\$900.000

### **Costos por contratación de profesionales**

Una vez que el modelo indique la cantidad de alumnos con probabilidad de deserción, la Universidad deberá contratar tutores que estén a cargo de monitorear el desempeño de los alumnos y apoyarlos en su proceso de enseñanza aprendizaje.

Se considerará que los tutores que estarán a cargo de estas labores serán estudiantes de la misma casa de estudio que tienen un desempeño académico destacado. El pago que realizará la Universidad por la contratación de estos servicios, será de \$280.000 por cada tutor, el cual podrá tener hasta 4 estudiantes a cargo. El valor mencionado corresponde a un tutor, que tiene 4 alumnos a cargo.

<b>Año</b>	<b>Cantidad de tutores</b>	<b>Alumnos intervenidos</b>	<b>Valor</b>
2012	6	26	\$1.680.000
2013	14	54	\$3.920.000
2014	21	82	\$5.880.000
2015	29	114	\$8.120.000
2016	37	148	\$10.360.000

2017	47	186	\$13.160.000
------	----	-----	--------------

### Flujo de caja

Para realizar el flujo de caja, la tasa de descuento que se utilizará es del 15% ya que es la que ocupa la Universidad para evaluar proyectos y el horizonte de tiempo será de 3 años ya que dado lo rápido de los avances tecnológicos, quedan muy rápidamente obsoleta la tecnología.

Para poder realizar esta evaluación se van a considerar 3 escenarios posibles:

- Optimista: Se considerará que, dadas las estrategias a implementar, se logrará retener a un 80% de los estudiantes
- Conservador: Se considerará que, dadas las estrategias a implementar, se logrará retener a un 50% de los estudiantes
- Pesimista: Se considerará que, dadas las estrategias a implementar, se logrará retener a un 20% de los estudiantes

Para poder realizar el cálculo, se realizó el siguiente supuesto, el cual es bastante conservador. Considerando los valores de la tabla 64 se calculó la diferencia en cuanto a dinero respecto a los años 2017 y 2016, es cual dio la siguiente cifra: \$166.721.200.

Se consideró que este valor sería constante para los tres años siguientes, lo cual es un supuesto bien conservador, considerando que la tendencia en los últimos años ha ido en aumento. Según lo anterior, tenemos lo siguiente:

Año	Ingreso que se deja de percibir
2019	\$902.779.014
2020	\$1.069.500.214
2021	\$1.236.221.414

Tabla 64: "Proyección de Ingresos que se dejan de percibir"

### Escenario Optimista

	Año 0	Año 1	Año 2	Año 3
<b>Ingresos</b>				
Ingresos por estudiantes recuperados		\$722.223.211	\$855.600.171	\$988.977.131
<b>Costos directos</b>				
Tutores de acompañamiento		-\$15.960.000	-\$18.760.000	-\$21.560.000
<b>Costos de Mantención</b>		<b>-\$900.000</b>	<b>-\$900.000</b>	<b>-\$900.000</b>
<b>Resultado operacional</b>		<b>\$705.363.211</b>	<b>\$835.940.171</b>	<b>\$966.517.131</b>
Gastos en Administración				

Capacitación		-\$20.000.000	-\$20.000.000	-\$20.000.000
<b>Resultado no operacional</b>		\$685.363.211	\$815.940.171	\$946.517.131
Inversión inicial	- \$5.581.500			
<b>Flujo de caja</b>	<b>- \$5.581.500</b>	<b>\$685.363.211</b>	<b>\$821.540.171</b>	<b>\$957.717.131</b>

### Escenario Conservador

	<b>Año 0</b>	<b>Año 1</b>	<b>Año 2</b>	<b>Año 3</b>
<b>Ingresos</b>				
Ingresos por estudiantes recuperados		\$451.389.507	\$534.750.107	\$618.110.707
Costos directos				
Tutores de acompañamiento		-\$15.960.000	-\$18.760.000	-\$21.560.000
<b>Costos de Mantención</b>		<b>-\$900.000</b>	<b>-\$900.000</b>	<b>-\$900.000</b>
<b>Resultado operacional</b>		<b>\$434.529.507</b>	<b>\$515.090.107</b>	<b>\$595.650.707</b>
Gastos en Administración				
Capacitación		-\$20.000.000	-\$20.000.000	-\$20.000.000
<b>Resultado no operacional</b>		<b>\$414.529.507</b>	<b>\$495.090.107</b>	<b>\$575.650.707</b>
Inversión inicial	-\$5.581.500			
<b>Flujo de caja</b>	<b>-\$5.581.500</b>	<b>\$414.529.507</b>	<b>\$495.090.107</b>	<b>\$575.650.707</b>

### Escenario pesimista

	<b>Año 0</b>	<b>Año 1</b>	<b>Año 2</b>	<b>Año 3</b>
<b>Ingresos</b>				
Ingresos por estudiantes recuperados		\$180.555.803	\$213.900.043	\$247.244.283
Costos directos				
Tutores de acompañamiento		-\$15.960.000	-\$18.760.000	-\$21.560.000
<b>Costos de Mantención</b>		<b>-\$900.000</b>	<b>-\$900.000</b>	<b>-\$900.000</b>
<b>Resultado operacional</b>		<b>\$163.695.803</b>	<b>\$194.240.043</b>	<b>\$224.784.283</b>
Gastos en Administración				
Capacitación		-\$20.000.000	-\$20.000.000	-\$20.000.000

Resultado no operacional		\$143.695.803	\$174.240.043	\$204.784.283
Inversión inicial	-\$5.581.500			
Flujo de caja	-\$5.581.500	\$143.695.803	\$174.240.043	\$204.784.283

Es posible observar que en cualquier escenario (optimista, pesimista y conservador) el proyecto es rentable, ya que en todos los casos analizados, la inversión inicial es bastante baja y es posible obtener buenos resultados en cuanto a alumnos recuperados, siendo esto, no solamente bueno para la Institución de Educación Superior, sino que también para los alumnos, ya que cumplen con el objetivo de obtener una carrera profesional con la cual podrán desempeñarse, les permite una movilidad social y no quedan endeudados con una carrera que no son capaces de terminar.

## CAPÍTULO 9: CONCLUSIONES

Al finalizar este proyecto de tesis, se pueden identificar diferentes líneas de aprendizaje y conclusiones.

Una de las principales dificultades en la etapa de preprocesamiento y entendimiento de los datos fue la falta de registros y la poca uniformidad que había al momento de ingresarlos. La cantidad de datos faltantes para las 4 últimas notas de cada asignatura, era alrededor del 90%, es por esta razón que se decidió trabajar solamente con las 3 primeras calificaciones, sin embargo, las notas con las cuales se decidió trabajar, tampoco contaban con el 100% de los registros.

Tal como se puede observar en la Tabla 23 de la página 49, los datos faltantes para estas asignaturas eran alrededor del 50% para cada una de ellas.

Los datos definidos como faltantes eran celdas que contenían datos “0” o “1”. Los datos ingresados como “0”, fueron considerados un error de tipeo y fueron reemplazados por “1”. Dicho lo anterior el 50% de los datos faltantes eran calificaciones con nota 1 y no necesariamente, eran alumnos que desertaban, solamente era información que no se había ingresado y que no se contaba con ella.

Para poder construir la base de datos con la información que faltaba se utilizó la siguiente técnica:

- Los alumnos que no tenían registro en ninguna de las evaluaciones de la asignatura, pero se encontraban dentro de la clase “NO DESERTA”, se utilizaron los estadísticos descriptivos de la muestra con los datos que si se tenían. Con esta información, se utilizó la media y la desviación estándar, las cuales fueron ingresadas como input para generar números aleatorios que respetaran esa distribución.
- Para los alumnos que “DESERTAN” y para los cuales tampoco se tenía registro de sus calificaciones, se aplicó la misma técnica descrita anteriormente.

Dado que existencia de tantas notas faltantes, es difícil saber, en el caso de los alumnos que desertan si la calificación registrada como “1” es porque no rindió la prueba o porque no contestó nada correcto.

Para los alumnos que obtienen calificación 1, es necesario investigar por factores por los cuales no dan las pruebas: ¿están trabajando?, ¿tienen problemas que les impide llegar a la hora?, etc. Se recomienda que, para comenzar a entender este fenómeno, se solicite una justificación, la cual sea la condición para dar la prueba siguiente.

Se hace necesario comenzar a registrar las inasistencias y monitorear su comportamiento de los alumnos a través de otros canales, tales como:

- Carnet de biblioteca, revisar si sacan material de estudio;
- Ingreso al portal de la Universidad donde se publica el material de la asignatura y monitorear con la frecuencia que entran y descargan material.



Ambas recomendaciones apuntan a evaluar cuan activos o inactivos están los alumnos, de modo de poder citarlos y conocer las razones por las cuales no tienen un comportamiento acorde a las exigencias que requiere la carrera.

Actualmente, al momento de ingresar a la carrera, se le hace una encuesta a los matriculados la cual se divide en 5 líneas de investigación principalmente, las cuales son: (1) Antecedentes personales y familiares del estudiante, (2) Hábitos de Estudio y Autoconfianza, (3) Elección de la carrera, (4) Elección de la Universidad, (6) Percepción de la Universidad. Respecto a lo anterior, se puede concluir lo siguiente:

### **Lecciones aprendidas**

- Lo anterior, tiene relación con los resultados obtenidos con el modelo de predicción, ya que si durante el primer semestre, tienen malas calificaciones en los ramos principales, probablemente sientan que no tienen las capacidades para continuar y que dado sus condiciones socioeconómicas es mejor retirarse y así evitar una deuda.
- Para poder comprender el fenómeno de la deserción al interior de la Universidad, no sólo es importante contar con la información de las calificaciones de las asignaturas más importantes a tiempo, sino que también es importante considerar el perfil del alumno que ingresa.

### **Trabajos futuros**

- Respecto al registro y calidad de la información, se sugiere contar con estándares definidos respecto a la Gobernanza de los datos, definir bien los roles y la seguridad de la información. Hoy en día cada departamento cuenta con reglas propias para ciertos registros, lo cual hace difícil realizar ciertas comparaciones al momento de requerirlo. Los datos son el activo corporativo más valioso, permiten diseñar estrategias y planificar acciones. La comprensión de las diferencias entre el gobierno de datos y su gestión, o entre ésta y la gestión de la información permite identificar brechas en sus enfoques y crear una base que impulsará la calidad de los datos, mejorando la capacidad para tomar decisiones bien informadas.
- Se propone crear un plan de nivelación a todos los alumnos que ingresan en primer año, ya que en general, los alumnos no tienen una buena base (en función de los datos analizados anteriormente); esto ayudaría a los docentes a tener claridad respecto al nivel del curso, pudiendo ajustar los contenidos, en relación a las necesidades de los alumnos.
- El modelo propuesto, logra identificar las notas que ponen en riesgo la permanencia de los alumnos en la Universidad. Por lo tanto, es importante contar con las notas a tiempo, con el objetivo de tomar medidas preventivas y asignar tutores a los alumnos que lo requieran.
- Se propone crear un plan de trabajo que ayude a los alumnos a “aprender a aprender”, a gestionar sus emociones, tales como tolerar la frustración, gestionar de manera correcta el tiempo y prioridades.

## **CAPÍTULO 10: BIBLIOGRAFÍA**

Barros V, O. (2008). Diseño Integrado de Negocios, Procesos y Aplicaciones TI. Santiago, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial.

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.

Betancourt, G. A. (2005). Las máquinas de soporte vectorial (svms). *Scientia et technica*, 1(27).

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2008). Análisis de los factores asociados a la deserción estudiantil en la Educación Superior: un estudio de caso Analysis of the Factors Associated with the Drop-out Rate of Students in Higher Education: a Case Study. *Revista de Educación*, 345, 255-280.

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2009). Deserción estudiantil universitaria: una aplicación de modelos de duración. *Lecturas de economía*, 60(60), 39-65.

Castro, J. A. F., & González, D. M. S. (1993). Redes neuronales: algoritmos, aplicaciones y técnicas de programación.

Díaz Peralta, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios pedagógicos (Valdivia)*, 34(2), 65-86.

Ethington, C. A. (1990). A psychological model of student persistence. *Research in higher Education*, 31(3), 279-293.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27-34.

Fernández, A. J. J. (1995). Análisis de regresión logística. Centro de Investigaciones Sociológicas (CIS).

Hax, A. C., & Wilde, D. L. (1999). The delta model: adaptive management for a changing world. *Sloan Management Review*, 40(2), 11.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004) *Introducción a la Minería de datos*, Editorial Pearson Educación SA, Madrid

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

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

## CAPÍTULO 11: ANEXOS

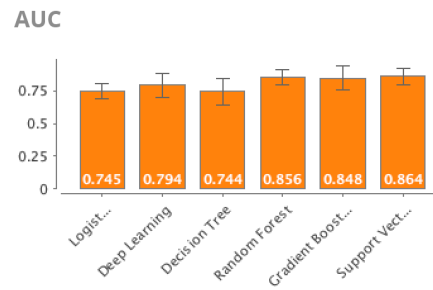


Gráfico 9: Comparación de AUC entre modelos

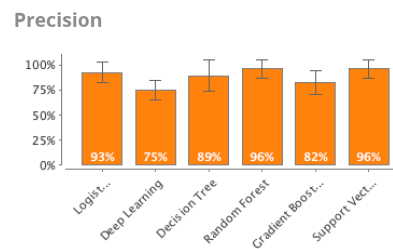


Gráfico 10: Comparación del indicador Precision entre modelos

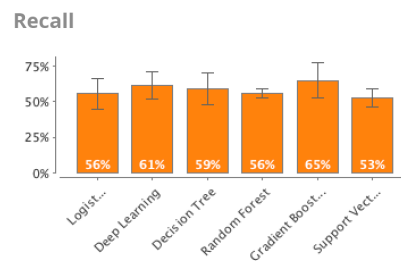


Gráfico 11: Comparación del indicador Recall entre modelos

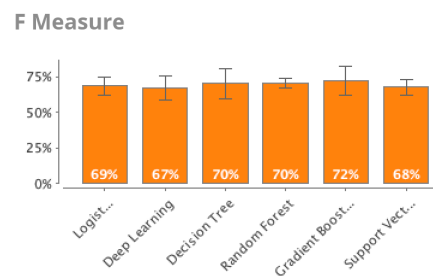


Gráfico 12: Comparación del indicador Fscore entre modelos

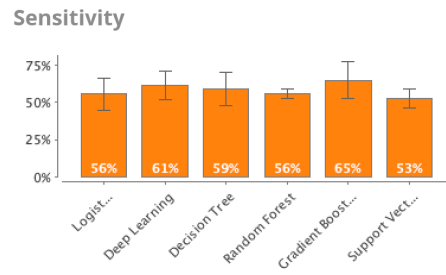


Gráfico 13: Comparación del indicador Sensitivity entre modelos

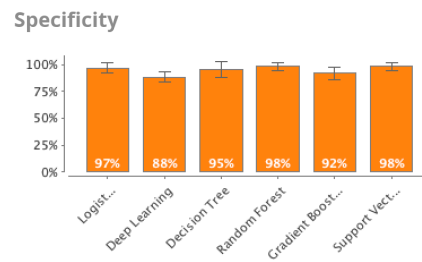


Gráfico 14: Comparación de indicador Specificity entre modelos