UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS

ESCUELA DE POSTGRADO

**PROGRAMA DE MAESTRÍA EN DIRECCIÓN DE SISTEMAS Y
TECNOLOGÍAS DE LA INFORMACIÓN**

**Propuesta de un Modelo Analítico de Datos para la Retención de Estudiantes
del Instituto Cibertec – Sede Norte**

TRABAJO DE INVESTIGACIÓN

**Para optar el grado académico de Maestro en Dirección de Sistemas y Tecnologías de la
Información**

AUTOR(ES)

Palacios Zavala, Ana Natalia (0000-0001-6494-0024)

Chavez Bardales, Víctor Isaac (0000-0003-0147-7321)

Lizarzaburu Li, Harold Vladimir (0000-0002-5562-9848)

Santos Navarro, Pascual Alexander (0000-0002-1834-3613)

ASESOR

Medina La Plata, Edison Humberto (0000-0001-5305-6516)

Lima, 07 de enero de 2020

DEDICATORIA

Quiero agradecer a mis padres, por los valores y perseverancia que me inculcaron y que son fundamentales en mi vida personal y profesional, también por ser un ejemplo ante la adversidad y por mantener firmes sus objetivos de vida, a mis hermanos en quienes busco ser un guía en su camino profesional y buscar contribuir en su bienestar personal.

A mi pareja Olga, por su apoyo incondicional, comprensión y aliento día a día y especialmente durante la maestría que me permitió culminar esta meta profesional.

Harold Vladimir Lizarzaburu Li

DEDICATORIA

Dedico de manera especial a mi madre Genny Navarro, por brindarme todo su apoyo y por ser el motor y motivo para siempre seguir adelante. Asimismo, a mi padre Pascual Santos por ser un ejemplo y no quebrantar sus valores ante cualquier eventualidad. A mis hermanos por siempre estar cuando más los he necesitado. A mi pareja Katherine por brindarme su apoyo incondicional.

Pascual Alexander Santos Navarro.

DEDICATORIA

A mis padres por la enseñanza de vida y por “el querer es poder”, a mi esposa Cristina por su apoyo incondicional y la fortaleza que siempre me brinda.

Victor Chavez Bardales.

DEDICATORIA

Agradezco a Dios por las oportunidades que me brinda, a mi familia por ser un soporte incondicional, a mis amigos. Un agradecimiento especial a mi equipo de trabajo universitario por las oportunidades.

Ana Natalia Palacios Zavala

RESUMEN

El estudio de la deserción a nivel nacional permite identificar los diversos motivos, razones y/o circunstancias por las que un estudiante decide cambiar de institución o, en el peor de los casos, no continuar sus estudios superiores. En los últimos años, este porcentaje de deserción se ha mantenido constante y no se han tomado medidas para identificarlo y/o tomar acciones para mejorar este escenario. Realizar una evaluación que permita identificar estudiantes con un alto riesgo de deserción en la institución y contar con un plan de acción debe ser parte del plan de continuidad estudiantil de las instituciones.

Con la aplicación de analytics y la minería de datos establecen un conjunto de actividades definidas en diferentes niveles de abstracción, las mismas que permiten analizar y estudiar los datos de manera estructurada con el fin de catalogar los registros, permitiendo clasificar la información de mayor a menor relevancia para analizar los patrones y lograr la identificación del perfil de un estudiante con alta probabilidad de deserción. Para la aplicación de estas técnicas de minería de datos se debe seguir las fases establecidas en metodologías, para ello se tomará como referencia la metodología CRISP-DM. Se recopiló información considerando los sistemas académicos y financieros de la institución, considerando las variables más relevantes para la construcción de los patrones y perfiles.

La presente tesis tiene por objetivo encontrar un patrón de comportamiento que nos permita definir perfiles de estudiantes, a través de la aplicación de analytics, lo cual permita identificar a tiempo el mayor grado de deserción y aplicar estrategias que permitan aumentar el nivel de retención de los estudiantes. Inclusive se analizarán aspectos económicos, que nos permiten demostrar el costo beneficio de la aplicación para la institución.

Palabras clave: Deserción; analytics; minería de datos; patrones; retención; perfiles de estudiantes.

**Proposed Analytical Data Model for the Retention of Students of the Cibertec
Institute - North Headquarters**

ABSTRACT

The research of the attrition in Peru let us seen different reasons and circumstances of why students change from schools or left their studies. In the last years, the percentage of attrition has remained constant, however, no actions have been taken. The evaluation of attrition, let us take the early actions to help students continue their studies.

The implementation of a solution based on analytics and data mining allows to have potencial data that can give us information to obtain profiles and patterns witch we can develop an strategy to avoid attrition. In order to devolop this solution, we need a methodology, in this case, we are using CRISP-DM methodology. In order to carry out the application of CRISP-DM, the collection of information is required considering the systems that provide relevant data for the construction of patterns and profiles.

The objective of this research is to find those patterns that can define a student profile, for this we are going to apply data analytics, after that we can be able to proposed some strategies to reduce attrition.

Keywords: data analytics, attrition, data mining, patterns and profiles

TABLA DE CONTENIDOS

1	INTRODUCCIÓN	1
2	MARCO TEORICO	3
2.1	SECTOR EDUCATIVO SUPERIOR	3
2.1.1	Institutos Superiores Tecnológicos	3
2.1.2	Institutos de Educación Superior (IES).....	3
2.1.3	Escuelas de Educación Superior (EES)	3
2.1.4	Acreditación de Institutos y Escuelas de Educación Superior	4
2.2	CONCEPTOS DE DESERCIÓN Y RETENCIÓN	6
2.2.1	Deserción estudiantil	6
2.2.2	Retención estudiantil	7
2.2.3	Deserción y Retención estudiantil en el contexto peruano.....	8
2.3	MODELOS APLICABLES EN EL ANÁLISIS	11
2.3.1	Modelos Predictivos.....	11
2.3.2	Modelos descriptivos	11
2.3.3	Modelos de toma de decisión	12
2.3.4	Modelos Ensemble	12
2.3.5	Modelo de aprendizaje supervisado.....	13
2.4	DATA ANALYTICS	18
2.4.1	Definición de analítica de datos.....	18
2.5	DATA MINING	18
2.5.1	Definición.....	18
2.5.2	Proceso Data Mining (DM).....	20
2.6	METODOLOGÍAS	21
2.6.1	Metodología fundamental para la ciencia de datos IBM	21
2.6.2	KDD (Knowledge Discovery in Databases).....	24
2.6.3	CRISP_DM (Cross Industry Standard Process for Data Mining)	25
2.6.4	SEMMA (Sample, Explore, Modify, Model, Assess).....	27
2.6.5	BASIM (Business Analytics Solutions Implementation Method)	29
2.7	HERRAMIENTAS PARA ANALÍTICA DE DATOS	30

3	SITUACIÓN ACTUAL Y SITUACIÓN PROBLEMÁTICA.....	35
3.1	SITUACIÓN ACTUAL	35
3.1.1	Objetivo de Estudio: El Instituto	35
3.1.2	Organización Administrativa:.....	36
3.1.3	Misión	42
3.1.4	Visión	42
3.1.5	Valores	42
3.1.6	Objetivos Generales de la institución.	42
3.1.7	Proceso Administrativos en Ciclo de vida del alumno	43
3.2	SITUACIÓN PROBLEMÁTICA	46
3.2.1	Descripción del Problema	46
3.2.2	Enunciado del problema	48
3.2.3	Justificación del problema	48
3.2.4	Alcance y limitaciones	49
4	SOLUCIÓN PROPUESTA.....	50
4.1	ANTECEDENTES.....	50
4.2	SELECCIÓN DE LA METODOLOGÍA	52
4.3	APLICACIÓN DE LA METODOLOGÍA	52
4.3.1	Comprensión del negocio	52
4.3.2	Comprensión de los datos	55
4.3.3	Preparación de los datos	71
4.3.4	Modelado	77
4.3.5	Evaluación	87
5	EVALUACIÓN ECONÓMICA	91
5.1	ANÁLISIS DE LA SITUACIÓN ACTUAL	91
5.2	PRESUPUESTO DEL PROYECTO	91
5.2.1	Consideraciones para la elaboración del presupuesto	91
5.2.2	Flujo de caja del Proyecto.....	93
5.2.3	Resultados en Beneficio para Cibertec	98
6	CONCLUSIONES	99
7	RECOMENDACIONES	100

8	GLOSARIO DE TERMINOS	102
9	REFERENCIAS	105

ÍNDICE DE TABLAS

Tabla 01: Criterio de Swets	17
Tabla 02: Etapas de la Ciencia de Datos	22
Tabla 03: Variables consideradas para el análisis de la deserción estudiantil.....	50
Tabla 04: Variables consideradas como factores de riesgo.....	52
Tabla 05: Plan de proyecto de analítica de datos.....	54
Tabla 06: Descripción de variables de análisis.....	58
Tabla 07: Resultado de auditoría de datos.....	70
Tabla 08: Resultado de calidad de datos	70
Tabla 09: Datos no relevantes	72
Tabla 10: Notas por curso.....	73
Tabla 11: Inasistencias por curso.....	73
Tabla 12: Agrupación de lugar de procedencia por zona	74
Tabla 13: Etiquetas de la variable Facultad.....	74
Tabla 14: Etiquetas de la variable productos (Carreras)	75
Tabla 15: Codificación de variables categóricas	76
Tabla 16: Tabla de clasificación.....	81
Tabla 17: Variables de la ecuación del modelo de Regresión Logística	82
Tabla 18: Extracto de predicción con el modelo de Regresión Logística	82
Tabla 19: Tabla de clasificación.....	84
Tabla 20: Extracto de predicción con el modelo CHAID	87
Tabla 21: Tabla de clasificación Regresión Logística (Validación).....	87
Tabla 22: Tabla de clasificación árbol CHAID (Validación).....	88
Tabla 23: Coeficientes de la Formula generada con el modelo de Regresión Logística.....	89
Tabla 24: Situación actual de la sede norte	91
Tabla 25: Presupuesto de Implementación.....	92
Tabla 26: Escenarios propuestos para el análisis económico	94
Tabla 27: Estado de Resultados Escenario Pesimista.....	94
Tabla 28: Evaluación económica Escenario Pesimista.....	95
Tabla 29: Estado de Resultados Escenario Moderado.....	96
Tabla 30: Evaluación económica Escenario Moderado	97

ÍNDICE DE FIGURAS

Figura 1: Relación de dimensiones y factores del modelo de acreditación.....	5
Figura 2: Deserción por Segmento Económico.....	9
Figura 3: Factores de Deserción.....	10
Figura 4: Fases de Modelo Descriptivo.....	11
Figura 5: Modelos de Tomas de Decisión.....	12
Figura 6: Algoritmo de Aprendizaje de Vlad Miller.....	13
Figura 7: Etapas del modelo de clasificación.....	14
Figura 8: Matriz de Confusión.....	16
Figura 9: Curva ROC.....	17
Figura 10: Clases de técnicas de minería de datos.....	19
Figura 11: Proceso de Minería de Datos.....	20
Figura 12: Metodología Fundamental para la Ciencia de Datos.....	22
Figura 13: Pasos del Proceso de Descubrimiento del Conocimiento (KDD).....	24
Figura 14: Esquema del proceso KDD resumido.....	25
Figura 15: Etapas del Proceso CRISP_DM.....	26
Figura 16: Etapas del Proceso SEMMA.....	28
Figura 17: Iteración de fases en SEMMA.....	29
Figura 18: IBM Metodo de Implementación de Solución de Analisis Empresarial.....	29
Figura 19: Entorno Gráfico WEKA.....	31
Figura 20: Entorno Gráfico SPSS.....	32
Figura 21: Entorno Gráfico R.....	33
Figura 22: Entorno Gráfico Orange.....	34
Figura 23: Sedes Cibertec.....	36
Figura 24: Organigrama Sede Administrativa.....	37
Figura 25: Organigrama de Sede.....	39
Figura 26: Líneas de negocio en CIBERTEC.....	40
Figura 27: Institutos según Calidad y Costo de los programas 2015.....	41
Figura 28: Costos de los Programas, 2015.....	41
Figura 29: Ciclo de vida del alumno de carreras técnicas.....	43
Figura 30: Etapa de captación del alumnos dentro del área Comercial.....	44
Figura 31: Alumnos matriculados por ciclo.....	45
Figura 32: Distribución de alumnos por sede Periodo 2019-I.....	45

Figura 33: Periodos académicos del ciclo 2019-I.....	46
Figura 34: Porcentaje de alumnos Desertores	47
Figura 35: Distribución de retiros por Sede.....	47
Figura 36: Porcentaje de retiros por Ciclo en Sede Norte	48
Figura 37: Estimación de pérdidas por ciclo en Periodo 2019-I, Sede Norte	49
Figura 38: Formulario de actualización de datos personales.....	55
Figura 39: Formulario de información de datos de alumnos.....	56
Figura 40: Formulario de información de notas actuales	56
Figura 41: Formulario de información de asistencia	57
Figura 42: Formulario de comprobantes de pago.....	57
Figura 43: Histograma de frecuencia para la variable PCT_INASISTENCIA_REAL.....	59
Figura 44: Histograma de frecuencia para la variable EDAD_ALUMNO	60
Figura 45: Histograma de frecuencia para la variable AVG_NOTA_REAL.....	60
Figura 46: Histograma de frecuencia para la variable NUM_CURSO	61
Figura 47: Histograma de frecuencia para la variable APROBADO_REAL	61
Figura 48: Distribución de alumnos por ciclo	62
Figura 49: Distribución de alumnos por carrera.....	63
Figura 50: Distribución de alumnos por facultad	64
Figura 51: Distribución de alumnos por condición de retiro.....	64
Figura 52: Distribución de alumnos por lugar de procedencia.....	65
Figura 53: Distribución de alumnos por tipo de ingreso	65
Figura 54: Diagrama de cajas Retiros por PCT_INASISTENCIA_REAL.....	66
Figura 55: Diagrama de cajas Retiros por EDAD_ALUMNO	67
Figura 56: Diagrama de cajas Retiros por AVG_NOTA_REAL	68
Figura 57: Diagrama de barras Retiros por SEXO_ALUMNO	68
Figura 58: Diagrama de barras Retiros por FACULTAD	69
Figura 59: Diagrama de barras Retiros por DES_PRODUCTO	69
Figura 60: Configuración de selección de características	71
Figura 61: Resultado de selector de características	72
Figura 62: Configuración de directiva de equilibrado de datos.....	76
Figura 63: Recuento de retiros antes de equilibrar datos.....	77
Figura 64: Recuento de retiros después de equilibrar los datos	77
Figura 65: Resultado del Clasificador Automático	78

Figura 66: Particionamiento de datos	78
Figura 67: Ruta “Modelo_desercion” en SPSS	79
Figura 68: Configuración del modelo de regresión logística	80
Figura 69: Importancia del predictor en modelo de Regresión Logística.	80
Figura 70: Resultados de precisión y Métricas en modelo de Regresión Logística	81
Figura 71: Configuración del árbol de decisión CHAID.....	83
Figura 72: Importancia del predictor en modelo CHAID.....	84
Figura 73: Resultados de precisión y Métricas en modelo CHAID	85
Figura 74: Resultados de precisión y Métricas en modelo CHAID	85
Figura 75: Árbol generado con el modelo CHAID	86
Figura 76: Comparación de resultados Regresión Logística y Árbol CHAID	88
Figura 77: Fórmula de regresión logística.....	89
Figura 78: Fórmula de regresión logística para la probabilidad de deserción.....	89
Figura 79: Fórmula implementada en Excel.....	90
Figura 80: Infraestructura propuesta.....	92

1 INTRODUCCIÓN

El presente trabajo tiene como reto estructurar toda la información del estudiante y transformarla en conocimiento. La analítica de datos es uno de los pilares de la transformación digital y a esto le sumamos la posibilidad de aprender de los datos.

Conocer la información histórica y obtener patrones que permitan estar un paso adelante de posibles situaciones y así generar valor en la organización aprendiendo de los datos, alcanzando los objetivos estratégicos de la institución. Considerando que cuando no se logre retener al estudiante, la institución perderá ingresos, afectándola económicamente.

El presente trabajo es una propuesta para el uso de la analítica para el estudio de los diversos patrones que permitan generar perfiles en los estudiantes con alta probabilidad de deserción, a fin de tomar decisiones y/o aplicar estrategias de manera efectiva, logrando reducir el número de alumnos desertores en la institución.

Para el logro del siguiente trabajo se ha estructurado en 7 capítulos: En el capítulo 1 se dará una breve introducción sobre lo que trata el trabajo. El capítulo 2 está compuesto por el marco teórico, donde se desarrollan los diversos conceptos como deserción, retención, instituciones públicas y privadas, normativas que rigen a las instituciones técnicas, el índice de deserción a nivel nacional. También se abarcan los conceptos de minería de datos, analítica, modelado y procesamiento de los datos, modelos analíticos de datos, las diversas metodologías existen para el procesamiento y análisis de los datos, herramientas para la ejecución de analítica.

El capítulo 3 plasma la situación problemática de la institución, Cibertec, que deseamos resolver con la ayuda de la analítica de datos, apuntando a los objetivos estratégicos de la institución. Se realizará un análisis de la situación actual de la sede norte, con respecto a la proporción de alumnos matriculados por periodo y el total de estudiantes que deserten.

El capítulo 4 muestra el desarrollo de la solución, poniendo en práctica los pasos de la metodología CRISP-DM, sobre los datos proporcionados por la institución, haciendo uso de la herramienta elegida, SPSS Analytics, y encontrar los valores óptimos para los patrones de deserción.

El capítulo 5, comprende el análisis económico de los resultados obtenidos luego de aplicado el modelo de retención de alumnos. Se aplicará el $P \times Q$ (costo por número de alumnos), es

decir sobre la cantidad de alumnos retenidos, considerando costos de aplicación de planes de retención.

El capítulo 6, comprende las conclusiones y recomendaciones en base a los resultados del análisis realizado en el capítulo 4 y 5.

En el capítulo 7, se encontrará todas las referencias bibliográficas utilizadas para el desarrollo del presente trabajo.

2 MARCO TEORICO

2.1 Sector Educativo Superior

2.1.1 Institutos Superiores Tecnológicos

El Congreso de la República indica en la (Ley N°30512, 2016, art. 1) que “El MINEDU regula la creación, licenciamiento, régimen académico, gestión, supervisión y fiscalización de los institutos de Educación Superior y Escuelas de Educación superior público y privados”. Los Institutos están en la segunda etapa del sistema educativo nacional con formación de carácter técnico, lo que proyecta la integración del conocimiento teórico e instrumental con la finalidad de realizar inserción laboral.

2.1.2 Institutos de Educación Superior (IES)

Los institutos de Educación Superior (IES) también pertenecen a la segunda etapa del sistema educativo nacional, con énfasis en una formación aplicada a nivel técnico. Los IES brindan formación debidamente fundamentada de carácter técnico. Además, brindan estudios de especialización de perfeccionamiento profesional en áreas específicas y otros programas de formación continua, y otorgan los respectivos certificados. La gestión de los IES públicos está a cargo del Organismo de Gestión de Institutos y Escuelas de Educación Superior Tecnológica Públicos (Educatec), creado por el artículo 43. Los IES otorgan el grado de bachiller técnico y los títulos de técnico y de profesional técnico a nombre de la Nación. Asimismo, pueden brindar a nombre de la Nación los certificados de estudios técnicos y de auxiliar técnico.

2.1.3 Escuelas de Educación Superior (EES)

Las escuelas de Educación Superior (EES) son instituciones educativas de la segunda etapa del sistema educativo nacional, que forman personas especializadas en los campos de la docencia, la ciencia y la tecnología, con énfasis en una formación aplicada, las EES brindan:

- a. Formación altamente especializada teórica, con aplicación de técnicas para resolver problemas o proponer nuevas soluciones.
Su ámbito de estudio y actuación es específico dentro de un área de conocimiento tecnológico, científico y de la docencia. Las EES desarrollan investigación y proyectos de innovación.
- b. Estudios de especialización, de perfeccionamiento profesional en áreas específicas y de otros programas de formación continua. Otorgan los respectivos certificados.

- c. Estudios de segunda especialidad y de profesionalización docente, según corresponda. Las EES vinculadas a la tecnología y a las ciencias aplicadas a los sectores productivos de la economía nacional se denominan Escuelas de Educación Superior Tecnológica (EEST). Brindan formación especializada con fundamentación científica y el desarrollo de la investigación aplicada. Se orientan fundamentalmente al dominio de las ciencias aplicadas; a la asimilación, desagregación, innovación, adaptación, mejoramiento y modificación de la tecnología. Las EES vinculadas a la pedagogía se denominan Escuelas de Educación Superior Pedagógica (EESP) como centros especializados en la formación inicial docente. Forman, en base a la investigación y práctica pedagógica, a los futuros profesores para la educación básica y coadyuvan a su desarrollo profesional en la formación continua. Brindan programas de formación pedagógica que responden a las políticas y demandas educativas del país. Las EES otorgan grado de bachiller, y título profesional a nombre de la Nación, que es válido para estudios de posgrado. Las EEST otorgan también el grado de bachiller técnico y los títulos de técnico y de profesional técnico. La gestión de las EEST públicas está a cargo de Educatec y la gestión de las EESP públicas está a cargo del Ministerio de Educación, en coordinación con los gobiernos regionales

2.1.4 Acreditación de Institutos y Escuelas de Educación Superior

El ministerio de Educación indica que según (Decreto supremo N° 010-2017, 2019) las condiciones básicas de calidad son requerimientos mínimos para la provisión del servicio educativo en las instituciones de Educación superior. Su cumplimiento es necesario para el licenciamiento del instituto y la escuela de educación superior, de sus programas de estudios y de sus finales. Contemplan los aspectos detallados en el artículo 25 de la ley y se desarrollan en la norma que emite el ministerio de educación, la que establece componentes, indicadores, medios de verificación y otros necesarios para su evaluación

La acreditación es un mecanismo para la mejora continua de los programas de estudio del nivel superior, es el reconocimiento público y temporal a la calidad demostrada. Es un proceso voluntario, excepto para las carreras de Educación y Salud. Para optar por una educación de calidad es necesario que los modelos educativos se encuentren acreditados.

Los diversos modelos de calidad para la acreditación forman parte de un proceso para que las instituciones tenga la oportunidad de analizar su realidad actual y en base a dicho análisis poder ejecutar cambios que permitan mejorar constantemente su calidad educativa la que

permita autorregularse e impartir una cultura de calidad en todos los ámbitos de su formación.

El fin de la evaluación es encontrar aquellos puntos que pueden ser fortalecidos a la vez que permita un análisis profundo de los diversos aspectos de la calidad educativa. Tienen cuatro dimensiones según el modelo de acreditación de SINEACE: los resultados basados en el soporte institucional, la formación integral,

Figura 1: RELACIÓN DE DIMENSIONES Y FACTORES DEL MODELO DE ACREDITACIÓN

FUENTE: SINEACE MODELO DE ACREDITACIÓN INSTITUCIONAL PARA INSTITUTOS, 2017

En la etapa de autoevaluación, es la propia carrera o institución la que se evalúa y compara sus resultados con los estándares de calidad, así determina su nivel de logro e implementa acciones de mejora. En la etapa de evaluación externa, una Comisión de Evaluación Externa verifica el logro de los estándares de calidad y emite la recomendación de acreditación. La Comisión de Evaluación Externa está conformada por evaluadores externos registrados por el Sineace. El Evaluador Externo es un profesional con amplia experiencia que conoce la metodología de Evaluación Externa. El evaluador externo es un par evaluador, en tanto comparte el lenguaje, los códigos disciplinarios y profesionales, así como el enfoque de la institución que visita para la evaluación externa.

- ✓ El Ministerio de Educación (Minedu) ha otorgado hasta el momento el licenciamiento a 24 institutos de educación superior. La Ley de Institutos contempla cinco condiciones básicas de calidad (CBC).

- ✓ Se exige una gestión institucional que sustente una solidez organizativa con congruencia cumpliendo con el modelo educativo que se propone.
- ✓ Se exige que el área académica y los programas de estudios estén alineadas a las políticas del Minedu
- ✓ Se debe cumplir con una Infraestructura, recursos y equipamiento para el aprendizaje apropiado como laboratorios, centro de estudios de información, etc.
- ✓ Se debe contar con el 20% debe ser docentes a tiempo completo.
- ✓ Previsión económica y financiera compatible con los fines.

2.2 Conceptos de Deserción y Retención

Deserción y retención estudiantil son términos que están profundamente relacionados, razón por la cual resulta conveniente analizarlos en conjunto. Sin embargo, para poder entender dicha relación es necesario definir muy bien lo que significa cada uno de ellos.

Enfocada en un panorama global, la deserción estudiantil es compleja e incluye diversas causas. Los resultados muestran diversos factores que motivan la deserción; los mismos que iremos sustentando con el fin de enfocarnos en estos puntos claves y revertir la situación con el fin de lograr una retención del estudiante en la institución.

2.2.1 Deserción estudiantil

Todos los ciudadanos tienen derecho a una educación con igualdad de oportunidades; pero vemos que muchos jóvenes fracasan y toman la decisión de abandonar sus estudios, esto merma de forma negativa en su persona y su futura integración con la sociedad.

El concepto de deserción requiere de una definición dentro del marco normativo, y se entiende como la interrupción o desvinculación del alumno dentro del proceso educativo, considerando como desertores a aquellas personas que tienen dos semestres de inactividad.

Vicent Tinto¹ (1989) comenta que: *“Desde el punto de vista individual desertar significa el fracaso para completar un determinado curso de acción o alcanzar una meta deseada, en pos de la cual el sujeto ingresó a una particular institución de educación superior, por consiguiente, la deserción no sólo depende de las intenciones individuales, sino también de*

¹ Vicent Tinto, profesor adjunto de Sociología y Educación en la Universidad de Columbia New York, Títulos: Licenciado en Sociología, Maestría en Sociología, Doctor en Filosofía, Especialización: Sociología de la Educación; Teoría de la Estratificación Social.

los procesos sociales e intelectuales a través de los cuales las personas elaboran metas deseadas en cierta universidad”

2.2.1.1 Tipos de deserción

Según lo que plantea (Fontalvo, Castillo, & Polo, 2014). “La deserción puede darse en los primeros ciclos académicos o en los últimos, aunque la tendencia es mayoritaria en los periodos iniciales”. Por otro lado, (Castaño, Gallón, Gómez, & Vásquez, 2004) y (González, 2005) indican que se dan dos tipos de abandonos respecto al tiempo y al espacio.

Tipos A:

- Deserción precoz: cuando el estudiante es admitido por la universidad y no se matricula.
- Deserción temprana: cuando el abandono se registra en los dos primeros años de la carrera.
- Deserción tardía: cuando el estudiante abandona los estudios a partir del quinto semestre en adelante.

Tipos B:

- Deserción interna: cuando el estudiante decide cambiarse de carrera en la misma institución de educación superior
- Deserción institucional: cuando el estudiante abandona la universidad para matricularse en otra.
- Deserción definitiva del sistema educativo: cuando el estudiante no se matricula en ninguna universidad

2.2.2 Retención estudiantil

La preocupación por la retención de estudiantes es un fenómeno más bien reciente, pese a que el fracaso estudiantil en la enseñanza ha sido consustancial a la existencia misma de las universidades e institutos, especialmente de algunas carreras profesionales. La visualización de la retención como problema es una temática en ciernes en nuestro medio, donde aún no se comprende a cabalidad su impacto social y económico, ni se entiende en toda su magnitud la red de implicancias que se derivan del fracaso en los estudios.

El análisis de la retención de estudiantes involucra un conjunto de dimensiones que atañen la eficiencia técnica de los procesos de formación, la eficiencia económica y el impacto social, aspectos claves para la gestión de las instituciones y su proyección como centros de formación.

Múltiples son los desafíos a los que se han debido enfrentar las instituciones de educación superior, siendo la retención de primer año uno de los factores más críticos en las diversas instituciones. Con la retención se busca diseñar e implementar modelos de retención, los mismos que se complementen con la instauración de un modelo educativo en todas las aulas, promoviendo una forma de trabajar con los estudiantes.

El objetivo de esta estrategia es impactar en el proceso de incorporación y promover la generación de un sentido de permanencia, atacando los frentes académicos como psicosociales.

2.2.3 Deserción y Retención estudiantil en el contexto peruano

Existen pocos trabajos de investigación sobre deserción. En un estudio realizado en el 2013, por la empresa de marketing Flanqueo², se halló que un 30% de los alumnos matriculados en universidades e institutos del Perú, se retiraban en el primer semestre, al 2019, este porcentaje no ha disminuido.

En un análisis por el tipo de orientación de la universidad, aquellas que se orientaban al:

- ✓ Segmento C un 28% de deserción.
- ✓ Segmento B un 21% de deserción.
- ✓ Segmento A un 15% de deserción.

En la Figura 2, muestra el segmento económico C, tiene menores posibilidades de continuar con estudios universitarios. Este porcentaje se mantiene para la deserción en los Institutos Técnicos, con un 31% en el segmento C, un 20% en el segmento B y un 13% en el segmento A.

Se puede apreciar que en el segmento C, la población es más propensa a retirarse de los estudios. Esto es debido a los factores económicos o académicos.

² Flanqueo, consultora desde el 2007 que realiza planes integrales de marketing y dirección comercial.

FIGURA 2: DESERCIÓN POR SEGMENTO ECONÓMICO

FUENTE: ELABORACIÓN PROPIA

Es importante señalar que las condiciones de deserción dentro del segmento C obedecen a otras motivaciones dentro de su entorno social y familiar, dado que existen otras preocupaciones que impactan en su decisión de ingresar a una institución educativa. Como se muestra en los estudios del INEI³, las razones principales son las económicas y laborales.

Se puede entender que, en comparación con otros países, la posibilidad de acceder a una universidad, es un fin para mejorar el status económico y social, y un paso natural después de la educación escolar, a diferencia de los estudios técnicos. Sin embargo, no se puede quitar crédito a los institutos dado que no dejan de ser contratados por algunos estudiantes. Es importante señalar que los bajos costos de matrícula de algunas universidades compiten con los costos de institutos técnicos y profesionales, distorsionando la calidad de los servicios educativos.

En la figura 3 se muestra los factores que conllevan a que los alumnos decidan desertar, mostrando que en mayor porcentaje corresponde a los aspectos económicos, con un 37.7%.

³ INEI: Instituto Nacional de Estadísticas e Informática

FIGURA 3: FACTORES DE DESERCIÓN

FUENTE: INEI- PRIMERA ENCUESTA NACIONAL DE LA JUVENTUD PERUANA DESERTORA, 2011

Actualmente, solo el 35% de jóvenes en el Perú tiene acceso a la educación superior. Sin embargo, “no es suficiente con que ellos ingresen a la universidad, sino que se debe velar por su permanencia y desarrollo profesional. De inicio a fin”.

“Las universidades y las organizaciones son sistemas interconectados que deben trabajar de manera interdisciplinaria para reducir las brechas. Así, hay que tener en cuenta la diversidad”⁴

⁴ Especialista Juan Carlos Guevara Bolaños. Escuela de Postgrado de la Universidad Continental

2.3 Modelos aplicables en el análisis

2.3.1 Modelos Predictivos

Se basan en datos históricos como actualizados, los cuales en base a cierta lógica van obteniendo resultados, decimos resultados en general, dado que pueden obtenerse puntos de riesgo como de mejoras.

2.3.2 Modelos descriptivos

Tienen como objetivo la búsqueda de datos para poder establecer tendencias, esto mediante la relación de los datos obtenidos, para ello se debe haber definido el problema con la finalidad de poder realizar las preguntas para la obtención de dichos datos. Tener en cuenta que esta metodología es cualitativa y se apoya de los pasos del método científico. En la figura 4 se muestran las fases que intervienen en este modelo.

FIGURA 4: FASES DE MODELO DESCRIPTIVO

FUENTE: ELABORACIÓN PROPIA

Si luego de realizar la revisión de los resultados, no se logra encontrar la relación que atiende las preguntas del problema, se deberá replantear lo seleccionado, retornando a la fase inicial del modelo.

2.3.3 Modelos de toma de decisión

Tal como el nombre lo indica, son modelos que te ayudan a tomar una decisión.

De acuerdo con lo elaborado por Fernando Lámbarry Vilchis, Luis Arturo Rivas Tovar, María del Pilar Peña Cruz ⁵, los modelos de toma de decisiones pueden ser individuales y organizacionales. En la figura 5 se muestran a los autores que han desarrollado las diferentes teorías y modelos.

FIGURA 5: MODELOS DE TOMAS DE DECISIÓN

FUENTE: LÁMBARRY, F.; RIVAS, L. A. Y PEÑA, M. P. , 2010

2.3.4 Modelos Ensemble

Conocidos como métodos combinados, esto mediante el uso de algoritmos de aprendizaje combinados, los cuales se definen como supervisados y no supervisados.

De acuerdo con Vlad Miller, en su artículo “*Explorando Algoritmos de Aprendizaje Automático Supervisado*” ⁶, se realiza la siguiente clasificación:

⁵ Lámbarry, F.; Rivas, L. A. y Peña, M. P. (2010), “Modelos de decisión bajo una perspectiva de análisis de sus procesos”. En Universidad y Empresa, 18: 146-173.

⁶ Vlad Miller, Ingeniero de Software, Desarrollador de software en Larnaca, Chipre - Miembro desde el 11 de noviembre de 2013

FIGURA 6: ALGORITMO DE APRENDIZAJE DE VLAD MILLER

FUENTE: FUENTE ELABORACIÓN PROPIA

La diferencia entre ambos radica, tal como sus mismos nombres lo indican, en que el primero, el supervisado, trabaja en base a las variables a las cuales se les va ingresando, en base a ello irá generando las etiquetas correspondientes a las variables. Mientras que, el segundo, no supervisado, deja que el algoritmo decida cuál será la estructura y encontrará patrones escondidos.

El aprendizaje profundo, está más orientado a redes neuronales.

2.3.5 Modelo de aprendizaje supervisado

2.3.5.1 Modelo de Clasificación

De acuerdo a (Haro Rivera, Zúñiga Lema, Vera Rojas, Meneses Freire, & Escudero Villa, 2018) la clasificación es una técnica supervisada, donde generalmente se tiene un atributo llamado “clase” y busca determinar si los atributos pertenecen o no a un determinado concepto. También manifiesta que es una habilidad para adquirir una función que mapee (clasifique) un elemento de dato a una de entre varias clases predefinidas.

El problema fundamental de la clasificación está directamente relacionado con la separabilidad de las clases.

Las etapas del modelo de clasificación son las siguientes:

i. Construcción del modelo

Es la etapa en la cual se describe el conjunto de clases (predeterminadas), se le conoce como fase de aprendizaje. Cada instancia (tupla) se sabe que pertenece a una clase (etiqueta del atributo clase).

Se utiliza un conjunto de instancias para la construcción del modelo al cual se le denomina conjunto de entrenamiento (training set)

El modelo obtenido se representa como un conjunto de reglas de clasificación, árboles de decisión, fórmula matemática, entre otros.

ii. Utilización del modelo

Corresponde a la etapa de validación.

Se realiza la estimación de la precisión del modelo. Se utiliza un conjunto de instancias, distinto de los utilizados para la construcción del modelo, al cual se le denomina conjunto de prueba (test set). Si el conjunto de test no fuese independiente del de entrenamiento ocurriría un proceso de sobreajuste (overfitting).

Para cada instancia de test se compara la clase determinada por el modelo con la clase real (conocida).

La ratio de precisión es el porcentaje de ejemplos de test que el modelo clasifica correctamente.

FIGURA 7: ETAPAS DEL MODELO DE CLASIFICACIÓN

FUENTE: ELABORACIÓN PROPIA

A continuación, se describen los siguientes algoritmos de clasificación:

a. Árboles de Decisión

De acuerdo (Haro Rivera, Zúñiga Lema, Vera Rojas, Meneses Freire, & Escudero Villa, 2018) los árboles de decisión son técnicas de clasificación supervisada que permiten determinar la decisión que se debe tomar siguiendo las condiciones a cumplir desde la raíz hasta alguna de sus hojas. La construcción del árbol de decisión se realiza partiendo el conjunto de datos en dos o más subconjuntos de observación, luego este subconjunto es nuevamente particionado usando el mismo algoritmo. Los algoritmos de árboles de decisión más utilizados son: ID3, C4.5 y CART.

b. K-NN o Nearest Neighbours (Vecinos más cercanos)

El método K-NN emplea la clasificación supervisada estimando la distancia de cierto número de muestras (K vecinos) a la muestra que se pretende clasificar, determinando su pertinencia a la clase de la que encuentre más vecinos etiquetados, considerando el criterio de mínima distancia. Esta técnica es válida solo para datos numéricos, no para clasificadores de textos.

c. CN2 Rule Induction

El algoritmo CN2 es una técnica de clasificación diseñada para la inducción eficiente de reglas sencillas y comprensibles de forma “si entonces”. Este modelo funciona solo para clasificar. La búsqueda de reglas puede ser por medida de evaluación, selecciona una regla heurística para evaluar las hipótesis encontradas; puede darse por: entropía (medida de la imprevisibilidad del contenido), mediante la precisión de Laplace o por una precisión relativa ponderada.

Evaluación de Modelo de aprendizaje supervisado

En el estudio de (Haro Rivera, Zúñiga Lema, Vera Rojas, Meneses Freire, & Escudero Villa, 2018) se manifiesta que para la evaluación de los algoritmos de clasificación se utilizan los siguientes indicadores:

- **Precisión y exactitud:** La precisión se refiere a la dispersión del conjunto de valores obtenidos de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión; la exactitud, también conocida como Accuracy, se refiere a cuán cerca del valor real se encuentra el valor medido. Para calcular estos indicadores se hace uso de la siguiente fórmula:

$$\text{Precisión} = \frac{tp}{tp + fp}$$

$$\text{Exactitud} = \frac{tp + tn}{tp + tn + fn + fp}$$

Donde:

tp: Verdadero Positivo

fp: Falso Positivo

tn: Verdadero Negativo

fn: Falso Negativo

- **Recall (sensibilidad o exhaustividad):** es la proporción de verdaderos positivos entre todos los valores positivos en los casos. Si su resultado es uno, entonces se han encontrado verdaderos positivos en la base de datos, por lo que no existiría ruido ni silencio informativo. Por el contrario, si su valor es cero los datos no poseen relevancia alguna. Se calcula por:

$$\text{Recall} = \frac{tp}{tp + fn}$$

Donde:

tp: Verdadero Positivo

fn: Falso Negativo

- **Matriz de confusión:** La matriz de confusión permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Las columnas representan el número de predicciones de cada clase, mientras que cada fila representa las instancias en la clase real. A continuación, se muestra la matriz de confusión para el caso donde se tienen dos clases

FIGURA 8: MATRIZ DE CONFUSIÓN

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	VN (verdadero negativo)	FP (falso positivo)
	Positivo	FN (falso negativo)	VP (verdadero positivo)

FUENTE: HARO RIBERA ET AL, 2018

- **Curva ROC:** De acuerdo a (Valle Benavides, 2017) la curva ROC es una herramienta estadística que se usa para el análisis de clasificar la capacidad discriminante de una prueba diagnóstica dicotómica. Se basa en una variable de decisión, cuyo objetivo es clasificar a los individuos de una población en dos grupos: uno que presente un evento de interés y otro que no.

FIGURA 9: CURVA ROC

FUENTE: DEL VALLE A., 2017

- **AUC (Área bajo la curva)**

El área bajo la curva es el estadístico por excelencia para medir la capacidad discriminante de la prueba. También permite comparar pruebas entre sí y determinar cuál es la más eficaz.

De acuerdo al criterio de Swets, se dicta que un área por debajo de 0,7 tiene una baja capacidad discriminante y que una por debajo de 0,9 puede ser útil para algunos propósitos, catalogando a las mayores de 0,9 con alta exactitud⁷.

TABLA 01: CRITERIO DE SWETS

Baja Exactitud	0,5 – 0,7
Útiles para algunos propósitos	0,7 – 0,9
Exactitud Alta	0,9 -1,0

FUENTE: VALLE BENAVIDES (2017)

⁷ Swets JA. (1988): 'Measuring the accuracy of diagnostic systems'. Science 240: 1.285-1.293.

2.4 Data Analytics

2.4.1 Definición de analítica de datos

De acuerdo con Analítica de Negocios⁸, podemos definir la analítica de datos predominantemente como una variedad de aplicaciones, desde inteligencia de negocios básica, elaboración de informes y procesamiento analítico en línea (OLAP) hasta diversas formas de análisis avanzados.

En ese sentido, es similar en naturaleza a la analítica de negocios, otro término general para los enfoques de análisis de datos, con la diferencia de que este último está orientado a los usos comerciales, mientras que el análisis de datos tiene un enfoque más amplio.

Las iniciativas de análisis de datos pueden ayudar a las empresas a aumentar los ingresos, mejorar la eficiencia operativa, optimizar las campañas de marketing y los esfuerzos de servicio al cliente, responder más rápidamente a las tendencias del mercado emergente y obtener una ventaja competitiva sobre los rivales, todo con el objetivo final de aumentar el rendimiento empresarial.

Dependiendo de la aplicación en particular, los datos que se analizan pueden consistir en registros históricos o información nueva que se ha procesado para usos de análisis en tiempo real. Asimismo, este proviene de una combinación de sistemas internos y fuentes de datos externas.

2.5 Data Mining

2.5.1 Definición

Existen varias formas de poder procesar datos y convertirlos en información. Una de ellas corresponde a la minería de datos, en inglés conocido como Data Mining. Esto conceptualmente refiere al proceso de extracción de conocimiento que es de interés para el usuario.

La minería de datos corresponde a un paso dentro de la metodología KDD, la cual consiste en técnicas de extracción y medidas de interés.

La minería de datos nace de la necesidad de conocer información útil a partir de los bases de datos o Data warehouse. Con el crecimiento de los datos disponibles, la inteligencia de negocios tuvo que dar paso a la aplicación de la minería de datos en soluciones empresariales

⁸ Analítica Negocios, empresa dedicada a Desarrollos informáticos y de sistemas orientados a los datos, la inteligencia de negocios y la analítica. <https://www.analiticanegocios.com/analitica-de-datos/>

y comerciales, puesto que de esta manera se permite el descubrimiento automático o semiautomático de información relevante a partir de estos cúmulos de datos.

En las ciencias y la ingeniería existe un amplio rango de problemas y dominios de aplicación para la minería de datos (Grossman, Kamath, Kegelmeyer, Kumar, & Namburu, 2013) . Se encuentran soluciones a partir de minería de datos para problemas de los campos de mercadeo, comercio, salud, predicción, transporte, meteorología, entre otros.

De acuerdo al estudio de (Gonzales Cam & Rodriguez Dominguez, 2017), las técnicas de minería de datos se distribuyen en predictivos, descriptivos y auxiliares. Dentro de los modelos que usan técnicas predictivas se tiene la regresión, series temporales, métodos bayesianos y de clasificación

FIGURA 10: CLASES DE TÉCNICAS DE MINERÍA DE DATOS

FUENTE: GONZALES Y RODRÍGUEZ, 2017

2.5.2 Proceso Data Mining (DM)

Considerando la denominación de proceso a Data Mining, se sobreentiende que hablaremos de fases en su ejecución. Para este proceso se han definido 5 fases las cuales vemos el siguiente gráfico:

FIGURA 11: PROCESO DE MINERÍA DE DATOS

FUENTE: ELABORACIÓN PROPIA

A continuación, procederemos a describir cada una de las fases del proceso DM:

a. Recolectar la data

Se debe considerar que en esta etapa se obtienen muchos datos de los cuales no todos, necesariamente, nos serán de utilidad. Es por ello que se debe iniciar una depuración y tomar los datos relevantes para el objetivo que se quiere alcanzar. En este punto tampoco se puede utilizar algoritmos complejos dato que la data no es confiable si no está depurada.

b. Pre-procesar la data

Esta fase ayudará a la eliminación de aquellos datos que no son los adecuados y no suman al cumplimiento del objetivo. También es posible utilizar las siguientes técnicas: redondeo, clustering, etc. En otras palabras, se tomarán las variables más “influyentes” con calidad.

Los métodos para la selección de características son básicamente dos:

- Aquellos basados en la elección de los mejores atributos del problema y aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

c. Estimar el modelo

Tal como se mencionó anteriormente, se utiliza el método descriptivo, generando modelos de asociación entre las variables, los cuales representan patrones de comportamiento. Las técnicas que se usan pueden variar dependiendo de las variables que se tienen.

d. Descifrar el modelo y extraer conclusiones

Al obtener el modelo, se tiene que validar, comprobando que las conclusiones que muestran son coherentes. Sin embargo, si se valida que este contiene varios modelos, se recomienda contrastar los modelos que mejor se ajuste al problema. y si el resultado no alcanza lo esperado con ninguno de los modelos se debe modificar los pasos anteriores para generar nuevos modelos.

2.6 Metodologías

Según (International Business Machines (IBM), 2015), la metodología es una estrategia general que orienta los procesos y las actividades dentro de un campo determinado. No hay dependencia de tecnologías o herramientas específicas, tampoco se trata de un conjunto de técnicas o fórmulas, por el contrario, una metodología proporciona un marco que determina la manera de proceder con los métodos, procesos y heurística que se utilizarán para obtener las respuestas o los resultados.

A continuación, se describen algunas metodologías relacionadas a la analítica de datos.

2.6.1 Metodología fundamental para la ciencia de datos IBM

De acuerdo a (International Business Machines (IBM), 2015), se requiere una metodología fundamental a medida que las capacidades de análisis de datos son más accesibles y frecuentes, esta metodología debe ser capaz de ofrecer una estrategia orientativa, al margen de las tecnologías que se puedan utilizar, los volúmenes de datos y los enfoques que intervengan. La metodología propuesta por IBM mantiene similitudes con otras ya conocidas para la minería de datos, pero enfatiza nuevas prácticas para la ciencia de datos, como por ejemplo el uso de grandes volúmenes de datos, la incorporación de análisis de texto para análisis predictivos y la automatización de algunos procesos.

Esta metodología consta de 10 etapas que se ejecutan de manera iterativa, tal como se muestra en la figura 12.

FIGURA 12: METODOLOGÍA FUNDAMENTAL PARA LA CIENCIA DE DATOS

FUENTE: IBM , 2015

Descripción de las Etapas

TABLA 02: ETAPAS DE LA CIENCIA DE DATOS

Etapa	Descripción
1: Comprensión del negocio	En esta etapa se define el problema que se necesita solucionar a nivel empresarial, además se establecen los objetivos del proyecto y los requisitos de la solución, siempre desde el punto de vista empresarial. Esta etapa es importante porque se establecen las bases para resolver con éxito el problema empresarial, por otro lado, la participación de los patrocinadores ayudará a garantizar el éxito del proyecto.
2: Enfoque Analítico	Se define un enfoque analítico para la solución del problema, expresándolo en un contexto de técnicas estadísticas y de aprendizaje automático. Es acá donde se puede definir qué modelo ayuda a la resolución del problema, por ejemplo, si el objetivo es predecir una respuesta como “sí” o “no”, podría definirse como la creación, prueba e implementación de un modelo de clasificación.
3: Requisitos de Datos	En la etapa anterior, con el enfoque analítico se determina los requisitos de datos, definiendo formatos, contenido y

	representaciones de datos, guiados por los conocimientos especializados.
4: Recopilación de Datos	En esta etapa se identifican y recopilan los datos que se encuentren disponibles y que sean relevantes para la naturaleza del problema, estos pueden ser estructurados, no estructurados y semiestructurados.
5: Comprensión de datos	Se hace uso de la estadística descriptiva y técnicas de visualización para comprender el contenido de los datos recopilados en la etapa anterior. Se puede revelar conocimientos iniciales sobre los datos y evaluar su calidad.
6: Preparación de datos	En esta etapa se tiene las actividades para construir el conjunto de datos que se usará en la etapa de modelado. Incluye la limpieza, combinación y transformación de datos. Generalmente es la fase más lenta del proyecto.
7: Modelado	Teniendo como base la primera versión del conjunto de datos, esta etapa se centra en el desarrollo de modelos predictivos o descriptivos conforme al enfoque analítico planteado en la etapa 2. Con los modelos predictivos, se utiliza un conjunto de datos de entrenamiento para la construcción del modelo. Esta etapa es altamente iterativa, permitiendo ajustar la preparación de datos y la especificación del modelo. Se puede probar varios algoritmos y seleccionar el mejor modelo.
8: Evaluación	Antes de la implementación del modelo, éste es evaluado para que se pueda comprender su calidad y pueda garantizar que se aborde de manera adecuada y completa el problema definido. Se hace uso de diferentes medidas de diagnóstico para interpretar la calidad del modelo y su eficacia para resolver el problema empresarial. En modelos predictivos, se hace uso de un conjunto de datos de prueba, el cual evalúa el modelo y realizar los ajustes si es necesario.
9: Implementación	Con el desarrollo del modelo satisfactorio y la aprobación de los patrocinadores empresariales, se procede a implementar en un entorno de producción. En este caso puede ser algo tan sencillo como la generación de un informe con recomendaciones o tan

	complicada como la integración del modelo en un flujo de trabajo complejo.
10: Retroalimentación	Con el modelo implementado se recopilan resultados para la retroalimentación del rendimiento y su impacto en el entorno. Esto permite ajustes al modelo para mejorar su precisión y utilidad.

FUENTE: ELABORACIÓN PROPIA

2.6.2 KDD (Knowledge Discovery in Databases)

KDD, por sus siglas en inglés es un término común en la minería de datos, es el descubrimiento de conocimiento en bases de datos (que viene a ser un proceso iterativo significativo), que consta de una serie de fases para la generación de conocimiento y la toma de decisiones. De acuerdo a (Hernandez, Ramirez, & Ferri, 2004) existe una diferencia entre estos dos términos, KDD es un proceso que consta de una serie etapas y minería de datos es parte de este proceso.

FIGURA 13: PASOS DEL PROCESO DE DESCUBRIMIENTO DEL CONOCIMIENTO (KDD)

FUENTE: ADAPTADO DE HAN ET AL, 2011

Las fases de KDD son:

1. Integración y recopilación. Consiste en establecer un entendimiento del dominio de la aplicación y de los conocimientos previos relevantes. En esta fase se determina también la selección de un conjunto de datos que pueden ser obtenidos de diferentes fuentes, sobre los cuales se realiza el descubrimiento.

2. Selección, limpieza y transformación. En esta etapa se seleccionan y preparan los datos que se van a minar. Sin embargo, existen factores como el ruido o valores atípicos que afectan la calidad de los datos, por lo que ante esta situación la limpieza es una de las tareas más importantes, puesto que permite la selección de la técnica que más se ajuste al problema a resolver.
3. Minería de datos. Es la fase más representativa, se determina qué tipo de tarea es la más apropiada, ya sea agrupamiento, reglas de asociación, correlación, clasificación, regresión, entre otras. Los resultados obtenidos dependen de las fases anteriores, por lo que existe la posibilidad de regresar a los pasos previos para requerir nuevos datos o para redefinir la solución al problema planteado.
4. Evaluación e interpretación. Los patrones descubiertos deben cumplir con tres propiedades: precisión, comprensibles e interesantes. En esta fase se evalúan e interpretan los patrones obtenidos. Algunas validaciones pueden ser a través de índices de evaluación, validación cruzada, matrices de confusión, entre otras.
5. Difusión y uso. Como última fase, el conocimiento descubierto debe de ser incorporado en algún sistema o simplemente documentarlo para su difusión a las partes interesadas. Este proceso incluye también la revisión y resolución de posibles conflictos con los conocimientos que anteriormente se tenía.

FIGURA 14: ESQUEMA DEL PROCESO KDD RESUMIDO

FUENTE: MOINE J. M., 2013

2.6.3 CRISP_DM (Cross Industry Standard Process for Data Mining)

Es una de las metodologías más utilizadas en proyectos de Data Mining (Aquino, Molero-Castillo, & Rojano, 2015), la cual establece un conjunto de actividades definidas en 4 niveles de abstracción (fases, tareas generales, tareas específicas e instancias del proceso),

estructurándose de forma jerárquica desde un nivel general hasta uno específico. En la figura 15 se muestran las etapas del nivel superior

FIGURA 15: ETAPAS DEL PROCESO CRISP_DM

FUENTE: IBM, 2012

1. Comprensión del negocio.

En este nivel se entienden los objetivos y requerimientos del proyecto desde la perspectiva del negocio con el fin de elaborar un plan preliminar que permita lograr estos objetivos.

2. Comprensión de los datos.

En esta etapa se inicia la recolección de datos para lograr familiarizarse con los mismos. Se identifica problemas con los datos como inconsistencias, errores, duplicidad, incompletos, etc.; esto permite iniciar acciones para asegurar la calidad.

3. Preparación de los datos.

Esta etapa abarca las actividades para la construcción del conjunto de datos que se va a utilizar, en estas se realiza la limpieza de datos, creación de variables adicionales, integración de datos y formato de trabajo.

4. Modelado.

Se inicia con la construcción del prototipo, haciendo uso de las técnicas apropiadas. La exploración de la diversidad de técnicas, que puedan solucionar el problema, puede llevar a retornar a la etapa anterior para un ajuste de datos.

5. Evaluación.

La importancia de esta etapa permite asegurar si el modelo obtenido logra alcanzar el objetivo inicialmente propuesto, con los resultados obtenidos se puede concluir en la aceptación del modelo. Tal como se muestra en la figura 14 se puede retornar a la revisión del objetivo en caso que los resultados no sean aceptados

6. Despliegue.

Luego de tener el modelo aprobado, éste es desplegado para uso de los usuarios finales. La manera de despliegue puede ser algo tan simple como la elaboración de un reporte, o tan compleja como la inclusión del modelo en el flujo de trabajo a través de una aplicación personalizada.

De acuerdo a lo indicado por (Rivo, y otros, 2012) esta metodología hace que los proyectos, grandes o pequeños, de minería de datos sean rápidos de desarrollar, fiables y manejables, gracias a su característica que hace énfasis en los detalles de cada fase, es decir, cada etapa se divide en diferentes tareas y actividades

2.6.4 SEMMA (Sample, Explore, Modify, Model, Assess)

De acuerdo a (Aquino, Molero-Castillo, & Rojano, 2015) SEMMA es una metodología creada por SAS Institute (Statistical Analysis Systems), el cual se define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones. Este proceso consta de cinco etapas (Figura 16) la cuales son necesarias para guiar el desarrollo de un proyecto de minería de datos (Moine J. M., 2013). (Sumathi & Sivanandam, 2006) mencionan que la metodología SEMMA permite aplicar la estadística exploratoria y técnicas de visualización de manera fácil, así como la selección y transformación de las variables más significativa, con el objetivo de crear modelos para predecir resultados y evaluarlos de manera que sirva de apoyo para la toma de decisiones

FIGURA 16: ETAPAS DEL PROCESO SEMMA

FUENTE: MOINE J. M. , 2013

1. Muestreo

Consiste en tomar una muestra de un conjunto de datos, que debe permitir obtener información relevante y además poder procesarse de manera de manera rápida, en este sentido, se debe tener la cantidad adecuada (ni tan grande, ni tan pequeña).

2. Exploración

Esta fase permite explorar los datos para familiarizarse con ellos y a la vez buscar tendencias y relaciones desconocidas. Se pueden formular nuevas hipótesis a partir de este análisis.

3. Modificación

Es una fase previa al modelado, en la cual se preparan los datos anómalos (valores atípicos y valores faltantes) para minimizar resultados erróneos, para ello se seleccionan, crean y modifican variables.

4. Modelado

En esta fase se crea el modelo que permitirá predecir variables, haciendo uso de técnicas y algoritmos predictivos como redes neurales, arboles de decisión, regresión o análisis discriminante.

5. Evaluación

A partir de los modelos obtenidos se evalúa la utilidad y exactitud que se han obtenido en el proceso de data Mining.

En la siguiente imagen se muestra la iteración de la metodología SEMMA

FIGURA 17: ITERACIÓN DE FASES EN SEMMA

FUENTE: MOINE J. M., 2013

2.6.5 BASIM (Business Analytics Solutions Implementation Method)

Metodología desarrollada por (IBM, IBM Business Analytics, 2011), la cual permite llevar a cabo un ciclo de vida de desarrollo completo para las soluciones de Business Analytics, por medio de una guía paso a paso. La metodología contiene roles y responsabilidades, pasos estructurados, pautas, actividades de desarrollo, prácticas comprobadas, consejos y documentos para solucionar problemas.

FIGURA 18: IBM METODO DE IMPLEMENTACIÓN DE SOLUCIÓN DE ANALISIS EMPRESARIAL

FUENTE: IBM BUSINESS ANALYTICS , 2011

1. Análisis

Se establece el objetivo de la solución a desarrollar, se incluyen características no funcionales (facilidad de uso, rendimiento, etc.). En esta etapa se debe obtener un acuerdo entre las partes interesadas acerca de estos requisitos.

2. Diseño

Los componentes de la solución son definidos, además de sus dependencias. En esta etapa permite identificar los recursos necesarios para el proyecto, además se trabaja en un ambiente de desarrollo.

3. Configuración y Construcción

En esta etapa se configuran, generan e integran los componentes utilizando un ciclo de desarrollo iterativo. Se hace uso del plan de validación.

4. Desplegar

En esta fase se elabora un plan para ejecutar y mantener la solución, incluyendo un programa de soporte. Se despliega la solución de analítica a un entorno de producción, se realizan las configuraciones según la necesidad y despliegue del plan de comunicación a los usuarios del negocio.

5. Mantenimiento

Representa el uso de la solución desarrollada e implementada. Esta etapa incluye tareas de mantenimiento y los puntos de control las cuales facilitaran el empleo exitoso de la solución y preservar su salud.

6. Gestión de Proyectos

Esta fase es Cross a las anteriores, consta de 4 procesos: Iniciar, Planear, Ejecutar y Cerrar, cada uno de los cuales ayuda a administrar y monitorear el mantenimiento del proyecto.

2.7 Herramientas para analítica de datos

2.7.1.1 Weka

Por sus siglas en inglés Waikato Environment for Knowledge Analysis, desarrollado en Java por la universidad Waikato en Nueva Zelanda en el año 1993, es una herramienta de software de minería de datos y aprendizaje automático. Este cuenta con algoritmos para la analítica de datos y modelamiento predictivo. A través de una interfaz gráfica permite unificar las herramientas de análisis. Su distribución es de licenciamiento GNU-GLP o software Libre.

Weka soporta muchas tareas estándares de data Mining, entre las cuales se identifican algoritmos de clasificación, regresión, clusterización, entre otros.

Dentro de las características que presentan esta herramienta se observa que funciona en múltiples plataformas, tiene una interfaz de usuario muy sencilla que permite el uso para usuarios inexpertos; por otro lado, permite el acceso a bases de datos por medio de SQL, a través del driver JDBC.

FIGURA 19: ENTORNO GRÁFICO WEKA

FUENTE: ELABORACIÓN PROPIA

2.7.1.2 SPSS Modeler

Originalmente nombrado como Clementine por sus creadores, luego, con la adquisición de SPSS por parte de IBM en 2009 pasó a llamarse IBM SPSS Modeler. Es una herramienta de software de minería de datos y análisis de texto la cual permite construir modelos predictivos y otras tareas analíticas. Permite que los usuarios aprovechen los algoritmos de estadística y de data Mining sin necesidad de programación, lo cual permite deshacer de la complejidad innecesaria en las transformaciones de datos y hacer que los modelos predictivos complejos sean muy fáciles de usar.

Esta herramienta es de pago teniendo las siguientes modalidades: IBM SPSS Modeler Personal, IBM SPSS Modeler Profesional, IBM SPSS Modeler Premium.

FIGURA 20: ENTORNO GRÁFICO SPSS

FUENTE: ELABORACIÓN PROPIA

2.7.1.3 R

Es una herramienta de código abierto, libre de licenciamiento bajo GNU.

Es uno de los lenguajes de programación más utilizados en la investigación científica y tiene una popularidad para realizar análisis de minería de datos integrando algoritmos de clasificación, agrupamiento, modelos lineales y no lineales, entre otros. Una de las características que se destacan es su capacidad gráfica. También puede integrarse a diferentes bases de datos.

Proporciona una amplia gama de técnicas y métodos para modelos predictivos, medios estadísticos, análisis y visualización de datos

FIGURA 21: ENTORNO GRÁFICO R

FUENTE: ELABORACIÓN PROPIA

2.7.1.4 Orange

Es una herramienta que contiene, de manera interactiva, componentes de flujo de trabajo para el análisis de datos. Fue creada en 1996 y tiene desarrollado el Machine Learning basado en framework de C++. Presenta algoritmos de aprendizaje y técnicas de proyección de datos, con widgets para transformación de datos, visualización, clasificación, regresión y evaluación. Contiene complementos desarrollados en Python y es un software de código abierto.

FIGURA 22: ENTORNO GRÁFICO ORANGE

FUENTE: ELABORACIÓN PROPIA

3 SITUACIÓN ACTUAL Y SITUACIÓN PROBLEMÁTICA

3.1 Situación Actual

3.1.1 Objetivo de Estudio: El Instituto

Cibertec es un instituto de educación superior tendiendo como líneas de negocios carreras técnicas, formación continua y Working Student en los ámbitos de tecnologías de información, Gestión, Diseño, Comunicaciones, Ingeniería y mecánica

Cibertec se fundó en el año 1983 con solo dos carreras técnicas: Computación e Informática y Electrónica. La corporación educativa Laureate International Universities, en el 2004, incorpora a Cibertec dentro de su flota de institutos y universidades. Fortaleciendo la credibilidad de la institución en formar profesionales integración con una educación de alta calidad.

En la actualidad, Cibertec cuenta con más de 18,000 estudiantes y está constituida por 6 sedes educativas en Lima Metropolitana, 1 sede en Arequipa y 1 sede en Trujillo. Como estrategia de expansión la red educativa Laureate International Universities incorporó en mayo del 2018 al instituto Avansys como parte de la red de institutos de Cibertec.

Esta situación posibilita que los estudiantes de las instituciones educativas amplíen las perspectivas de sus carreras y puedan, al mismo tiempo, acceder a alguna de las otras instituciones de Educación Superior de la Red. Sus miembros comparten los conocimientos y las mejores prácticas en el campo de la educación superior con el objetivo de formar profesionales preparados a nivel internacional. Cibertec, en los últimos años ha tenido un crecimiento considerable teniendo presencia en Cercado de Lima, Callao, San Juan de Lurigancho, Breña, Miraflores, Independencia y en provincia en Arequipa y Trujillo.

FIGURA 23: SEDES CIBERTEC

FUENTE: CIBERTEC, 2019

3.1.2 Organización Administrativa:

El IES Cibertec, cumple la estructura organizativa y las siguientes funciones según el artículo 76° del reglamento institucional.

FIGURA 24: ORGANIGRAMA SEDE ADMINISTRATIVA

FUENTE: REGLAMENTO INSTITUCIONAL CIBERTEC, 2018

3.1.2.1 Órganos de Dirección:

- **La Gerencia General**

Es la primera autoridad siendo responsable del planeamiento y dirección en la organización, estableciendo los objetivos, misión, visión y propósito. Manteniendo los estándares operativos que permitan la sostenibilidad del Instituto en el largo plazo.

- **La dirección Académica**

Es la mayor autoridad de nivel académico. Asimismo, este es el representante legal de la institución ante las direcciones regionales y el ministerio de educación. El director tiene como objetivo la planificación académica, establecer los objetivos del instituto y obtener resultados que contribuyan a la formación de los alumnos.

- **Dirección Comercial**

Responsable de analizar y desarrollar estrategias de captación de alumnos, esta dirección incluye el departamento del área de telemarketing de carreras técnicas, cursos cortos y marketing.

- **Gerencia Financiera**

Responsable de las normas y estrategias financieras en la organización. Con el equipo financiero se encargan de analizar el presupuesto para el logro de los objetivos y metas previstas.

- **Gerencia de Infraestructura**

Es el responsable de la implementación de las nuevas sedes así como también del mantenimiento de infraestructura del mismo cumpliendo los estándares que exige INDECI, instituto nacional de defensa Civil. Asimismo, se encarga de realizar las negociaciones con los contratistas en la ejecución de los proyectos y

controlando todo el proceso del proyecto hasta la entrega de la fecha comprometida del proyecto.

- **Gerencia de Talento Humano**

Es responsable de la dimensión de recurso humano en la institución basado en el reclutamiento, capacitación y contrato del mismo. Asimismo, proporciona mecanismos y lugares adecuados que propaguen la motivación y satisfacción en el trabajo.

3.1.2.2 Órganos de la línea:

- **Dirección de Calidad Educativa:**

Es encargada de diseñar e implementar estrategias y servicios para los profesores, garantizando la educación integral de los estudiantes con altos niveles de calidad.

- **Dirección de Servicio Educativos**

Encargado de cumplir con los estándares y políticas que exige la Minedu, Ministerio de Educación, y propiciar la educación con alta calidad.

La dirección tiene bajo su cargo las siguientes áreas:

- **Secretaria Académica**

Área encargada del control de asistencia de docentes, pago de planilla, atención a los estudiantes, registro de notas, coordinaciones y trámites con el Ministerio, registro de programación de clases, matrícula, proceso de admisión y emisión de constancias y certificados.

- **Aula y Laboratorios**

Área encargada de preparar todo el equipamiento y software para el desarrollo de las clases, Asimismo brinda el soporte de atención ante cualquier eventualidad en el desarrollo de clases.

- **Centro de Información**

Área encargada de brindar orientación al alumnado para la búsqueda y préstamo de libros digital o físicos.

- **Oportunidades Profesionales**

Área encargada de la inserción laboral de los alumnos de los últimos tres ciclos y alumnos egresados, con la finalidad de mantener la empleabilidad de los beneficiados.

FIGURA 25: ORGANIGRAMA DE SEDE

FUENTE: REGLAMENTO INSTITUCIONAL CIBERTEC, 2018

Oferta Educativa y Servicios

La institución además de poseer la línea de Carreras Técnicas, también cuenta con cursos de Extensión Profesional que es una línea que permite complementar a profesionales a través de cursos especializados y certificaciones de marcas reconocidas en el mundo de tecnología; finalmente una línea de Working Student que permite a personas que ya laboran y cuentan con cierta experiencia profesional previa, seguir una carrera. La figura 26 muestra las líneas de negocio que cuenta CIBERTEC actualmente.

FIGURA 26: LÍNEAS DE NEGOCIO EN CIBERTEC

FUENTE: ELABORACIÓN PROPIA

La línea de negocio de mayor valor para la institución corresponde a la de Carreras técnicas, por lo que el análisis del presente trabajo se enfoca en este rubro del negocio.

De acuerdo a (PRONABEC, 2015), CIBERTEC se encuentra en el cuadrante 2 de Institutos Superiores Tecnológicos (IST), que corresponden a de mayor calidad y alto costo, en la figura 27 se muestra la ubicación que posee la institución.

FIGURA 27: INSTITUTOS SEGÚN CALIDAD Y COSTO DE LOS PROGRAMAS 2015

FUENTE: PRONABEC, 2015

Cibertec puede llegar tener un costo de las carreras de S/. 21,924 en promedio y S/. 7,308 como costo anual (3 años de estudios, en promedio, de la línea de Carreras Técnicas), tal como se muestra en el cuadro de costos de programas.

FIGURA 28: COSTOS DE LOS PROGRAMAS, 2015

INSTITUTO	COSTO ANUAL			COSTO TOTAL DE LA CARRERA		
	PROM.	MÍN.	MÁX.	PROM.	MÍN.	MÁX.
TECSUP 1 -Sede Lima	S/. 17,796	-	-	S/. 53,388	-	-
Le Cordon Bleu	S/. 15,579	S/. 10,615	S/. 20,543	S/. 46,736	S/. 31,845	S/. 61,628
Chio Lecca	S/. 13,100	-	-	S/. 39,300	-	-
TECSUP 2 -Arequipa	S/. 9,300	-	-	S/. 27,900	-	-
Valle Grande	S/. 9,300	-	-	S/. 27,900	-	-
TECSUP 1-Trujillo	S/. 8,450	-	-	S/. 25,350	-	-
ADEX	S/. 8,320	-	-	S/. 24,960	-	-
CIBERTEC	S/. 7,308	-	-	S/. 21,924	-	-
Eurodiomas	S/. 7,300	-	-	S/. 21,900	-	-
IBEROTEC	S/. 6,900	-	-	S/. 20,700	-	-
Ceturgh Perú	S/. 6,300	S/. 4,800	S/. 7,800	S/. 18,900	S/. 14,400	S/. 23,400
Fundación DI Nuevo Pachacútec	S/. 5,397	S/. 5,397	S/. 5,398	S/. 16,192	S/. 16,190	S/. 16,195
IESTP "De Emprendedores - Lima" - USIL S.A.	S/. 4,930	-	-	S/. 14,790	-	-
Parroquia Espiritu Santo - IEST Trentino Juan Pablo II	S/. 4,900	-	-	S/. 14,700	-	-
Toulouse Lautrec	S/. 4,470	-	-	S/. 13,410	-	-
Sencico Lima	S/. 4,133	S/. 3,950	S/. 4,300	S/. 12,400	S/. 11,850	S/. 12,900
Parroquia Espiritu Santo - Trentino Juan Pablo II CETPROP P. Jose Luis Idigora Goya (*)	S/. 4,040	-	-	S/. 12,120	-	-
AVANSYS	S/. 3,943	S/. 3,400	S/. 4,150	S/. 11,828	S/. 10,200	S/. 12,450
Instituto de Formación Bancaria	S/. 3,700	-	-	S/. 11,100	-	-
Continental	S/. 3,193	S/. 2,600	S/. 3,820	S/. 9,578	S/. 7,800	S/. 11,460
Wernher Von Brawn	S/. 3,188	S/. 3,188	S/. 3,188	S/. 9,564	S/. 9,564	S/. 9,564
Sencico Arequipa	S/. 3,120	S/. 2,920	S/. 3,320	S/. 9,360	S/. 8,760	S/. 9,960
Corporación Khipu	S/. 2,990	S/. 2,840	S/. 3,140	S/. 8,970	S/. 8,520	S/. 9,420
SENATI	S/. 2,762	S/. 1,380	S/. 7,615	S/. 8,285	S/. 4,140	S/. 22,845

FUENTE: PRONABEC, 2015

3.1.3 Misión

Formar profesionales íntegros y competentes brindando una educación superior de alta calidad que contribuya al desarrollo del país.

3.1.4 Visión

Es ser la institución líder de educación superior técnica en el Perú con alcance a nivel nacional.

3.1.5 Valores

- **Integridad**

Comportarse y expresarse siempre con la verdad.

- **Respeto**

Consideración y reconocimiento de la dignidad de las personas y la integridad de la empresa.

- **Compromiso**

Promesa de cumplimiento de una responsabilidad o rol asignado.

- **Trabajo En Equipo**

Obrar en unión de otros con un mismo fin.

- **Pasión por la calidad**

Afición vehemente hacia la excelencia de nuestros productos, servicios y procesos.

3.1.6 Objetivos Generales de la institución.

- Desarrollar competencias en un entorno de aprendizaje real y seguro a través de proyectos y simulaciones de casos reales.
- Desarrollar proyectos con el enfoque de responsabilidad social a través de los lineamientos de la investigación aplicada.
- Potenciar la virtualización de los cursos de todas las carreras en las diferentes Escuelas Académicas.
- Promover el aprendizaje significativo, activo y experimental mediante la técnica de la gamificación relacionado con los intereses de los estudiantes.
- Ampliar la oferta educativa de los programas de estudio a través de otros sistemas de aprendizaje (sistema de educación dual, semipresencial y virtual).

3.1.7 Proceso Administrativos en Ciclo de vida del alumno

El cliente pasa por diferentes áreas para iniciar su proceso dentro de la institución clasificándose por diferentes estados tales como postulante, ingresante, alumno y egresado. Las áreas son de contacto directo con el alumno por lo que están supervisadas por el área comercial y servicios educativos.

FIGURA 29: CICLO DE VIDA DEL ALUMNO DE CARRERAS TÉCNICAS

FUENTE: ELABORACIÓN PROPIA

a. Captación de Clientes – Postulantes

El área comercial inicia diversas actividades para captar a futuros interesados en las carreras técnicas que imparte la institución a través de campañas comerciales en los distintos medios tales como radio, televisión, página web y redes sociales como Facebook y twitter. Asimismo, se potencia la base de alumnos con las visitas a los colegios generando mayor margen de captación con las visitas programadas a la institución. Los datos son registrados al sistema interno de la institución y este interactúa con el CRM y contact center para garantizar la comunicación de los clientes potenciales.

FIGURA 30: ETAPA DE CAPTACIÓN DEL ALUMNOS DENTRO DEL ÁREA COMERCIAL

FUENTE: ELABORACIÓN PROPIA

En este proceso los postulantes no son filtrados por un examen de admisión, que permita medir sus conocimientos. Por otro lado, no hay ninguna evaluación de test vocacional para saber si realmente el alumno ha escogido la carrera correcta. Estas dos evaluaciones son importantes ya que permitirá reducir el riesgo de deserción del alumnado durante el ciclo académico.

b. Etapa de Ingreso – Ingresante

En este proceso el ingresante entrega el certificado de estudios de secundaria, copia de DNI, 02 fotos tamaño carné, registros de apoderados y domicilio en la intranet de la institución, con la finalidad de generar la matrícula y la primera boleta de pago. En esta etapa no existe un seguimiento continuo al ingresante y por esta razón se han detectado retiros antes de iniciar las clases y lo denominaremos ingresantes con deserción precoz o “No Show”.

c. Etapa de Matrícula – Alumnos

En esta etapa el alumno ya cuenta con un código de alumno y es usado para las diferentes plataformas de la institución tales como intranet, aula virtual y correo electrónico que será usada durante toda la vida académica, en los seis ciclos de enseñanza. Como se puede apreciar en la figura 31, desde el año 2014 al 2019, la institución ha incrementado en un 50% de sus alumnos matriculados.

FIGURA 31: ALUMNOS MATRICULADOS POR CICLO

FUENTE: ELABORACIÓN PROPIA

Este incremento se debe a que los últimos años, desde el 2014, la institución ha tenido un crecimiento de sedes a nivel local en el Callao sede Bellavista, Lima Metropolitana sede Breña, San Juan de Lurigancho y a nivel nacional sede Arequipa, teniendo en la actualidad 7 sedes académicas. De acuerdo a la revisión actual, según la distribución de matriculados, en el periodo 2019 – I, podemos verificar que la mayor población de alumnos se encuentra en la sede norte (ubicado en el distrito de Independencia), con un 30% de la población total.

FIGURA 32: DISTRIBUCIÓN DE ALUMNOS POR SEDE PERIODO 2019-I

FUENTE: ELABORACIÓN PROPIA

La responsabilidad de proceso de matrícula de alumnos es de Secretaria Académica y está soportada por diferentes áreas tales como admisión, encargada de los registros de alumnos nuevos; el área de sistemas, responsable de la habilitación del servicio web para la matricula online. El área de coordinación académica, encargada de

programar los horarios de los ambientes y asignaciones de docentes para el registro de horarios de los alumnos.

d. Etapa de Conclusión de Estudios – Egresados

Los alumnos que han culminado su carrera se categorizan como egresados y pueden seguir un curso de diplomado para la emisión de Título profesional técnico. Los egresados pueden ingresar a la bolsa de trabajo de la institución, inscribiéndose con el personal de oportunidades profesionales, área responsable de insertar al alumno al entorno laboral. Si bien es cierto, al egresar un alumno no se considera amenaza para la deserción es importante que sus datos se mantengan actualizados en todos los sistemas.

3.2 Situación Problemática

3.2.1 Descripción del Problema

En la actualidad Cibertec llega a una ratio de deserción total del 33.64% (periodo 2019 I), lo cual ocasiona pérdidas económicas significativas. Esta situación hace que, a pesar del incremento de ingresantes en los últimos años, no se tenga resultados financieros que reflejen el crecimiento de la institución.

En Cibertec se desarrollan 02 periodos por año académico, iniciando el primero de ellos en el mes de marzo hasta julio y el segundo periodo en el mes de agosto hasta diciembre. En cada periodo se realiza un proceso de matrícula.

FIGURA 33: PERIODOS ACADÉMICOS DEL CICLO 2019-I

FUENTE: ELABORACIÓN PROPIA

La deserción en Cibertec se divide en 3 tipos que se muestran a continuación:

- **Deserción Inter-ciclo:** Son aquellos alumnos que culminaron el periodo anterior (2018-2) y no se matricularon en el periodo actual (2019-1).
- **Deserción Precoz:** Son aquellos alumnos potenciales que pagan su matrícula y no inician clases en el periodo actual (2019-1).

- **Deserción Regular “Retiros”**: son aquellos alumnos que inician clases en el periodo actual (2019-1) y en este periodo se retiran en cualquier momento.

El estudio se enfocará en la Deserción Regular “Retiros” por ser el grupo del cual se tiene registros de información para el análisis con un 11%. En la figura 34 podemos ver distribuido en sede norte los porcentajes según su tipo.

FIGURA 34: PORCENTAJE DE ALUMNOS DESERTORES

FUENTE: ELABORACIÓN PROPIA

En la figura 35 se muestra la distribución de retiros por sede en el periodo 2019-1. Tal como se puede apreciar, la sede Arequipa presenta el mayor porcentaje de retiros con un 43 % teniendo un total de alumnos retirados de 486 alumnos. Sin embargo, la más representativa en cantidad de alumnos en todo Cibertec es la de sede Norte con 1866 alumnos desertores. Siendo esta la más crítica ya que representa mayores pérdidas económicas.

FIGURA 35: DISTRIBUCIÓN DE RETIROS POR SEDE

FUENTE: ELABORACIÓN PROPIA

En la figura 36 se tiene el porcentaje de retiros por periodo para la sede Norte, se puede verificar que desde el 2017 I al 2019 I el porcentaje no desciende.

FIGURA 36: PORCENTAJE DE RETIROS POR CICLO EN SEDE NORTE

FUENTE: ELABORACIÓN PROPIA

3.2.2 Enunciado del problema

La deserción de estudiantes del instituto superior tecnológico CIBERTEC, genera pérdidas económicas de 4.58 Millones de soles Anuales.

3.2.3 Justificación del problema

Proyectando una estimación monetaria para la sede Norte, con el cálculo de alumnos retirados, los ciclos que aún faltan por concluir y el precio anual promedio que se paga por la carrera, CIBERTEC puede llegar a dejar de percibir 12.8 millones de soles, solo en una sede y en un periodo de análisis, esto se puede dar en mayor probabilidad si no se cuentan con estrategias de retención, basados en análisis predictivos, o mecanismos de reincorporación de alumnos, a corto plazo. En la figura 37 se realiza la estimación de perdidas basada en el porcentaje de retiros que se tienen por ciclo (para el periodo 2019 I), en el caso de alumnos retirados del primer ciclo, con un 78 % del total de retirados, se puede llegar a dejar de percibir más de 10 millones de soles.

FUENTE: ELABORACIÓN PROPIA

3.2.4 Alcance y limitaciones

En el presente estudio aplicará solo para la sede Norte, por ser la que tiene mayor participación de alumnos en CIBERTEC y representa la mayor cantidad de alumnos retirados. Para las muestras de datos y teniendo en cuenta las consideraciones anteriormente descritas, se realizará el análisis con datos del periodo 2019-I. No se consideran para el análisis a los alumnos desertores inter ciclo, ni precoz.

Los datos se obtendrán de los sistemas Socrates (gestión de notas y asistencia) y de Spring (sistema de gestión de pagos). Por lo que los resultados que se obtendrán tendrán relevancia académica y económica. No se ha encontrado un repositorio de datos relevantes para las condiciones sociales o culturales de los alumnos.

El análisis en el presente estudio busca identificar patrones los cuales nos permita evaluar las variables más determinantes para que la institución pueda focalizar sus estrategias de retención con el fin de lograr la permanencia de aquellos estudiantes con una alta probabilidad de desertar y de este modo la inversión que realice la institución en estas estrategias se vea refleja en la rentabilidad que representa la permanencia y conclusión de la formación profesional de sus estudiantes.

Así mismo como una limitación se tiene que la institución en la actualidad no contempla la implementación de servicios de información en la nube por temas estrictamente institucionales.

4 SOLUCIÓN PROPUESTA

4.1 Antecedentes

El artículo de (Barreno Vereau, 2012) muestra una metodología para comparar los modelos de “regresión logística” y “árbol de clasificación”, el cual se basó en el estudio de la deserción de una universidad particular. Los modelos usados obtuvieron un porcentaje de clasificación exitosa de por lo menos 94% en el análisis de la deserción universitaria.

El estudio hizo uso de las siguientes variables explicativas, las cuales fueron catalogadas como Académicas y Económicas. En la tabla 03 se muestra las variables usadas:

TABLA 03: VARIABLES CONSIDERADAS PARA EL ANÁLISIS DE LA DESERCIÓN ESTUDIANTIL

Factor	Variable	Valores de la variable
Académico	Efectividad Examen de Admisión (EEA). $EEA = \frac{\text{Puntaje obtenido Ex. Admisión}}{\text{Máximo puntaje posible}}$	[0;1]
Académico	Promedio Ponderado Acumulado (PPA). Media ponderada, según valor de créditos académicos de cada asignatura, de todas las notas obtenidas por el estudiante en las asignaturas que haya cursado hasta el periodo de análisis.	[0;20]
Académico	Proporción de créditos aprobados (PCA). $PCA = \frac{\text{Total créditos aprobados}}{\text{Total créditos matriculados}}$	[0;1]
Académico	Presenta acta de compromiso (AC). El acta de compromiso se extiende a aquellos alumnos con múltiples repitencias en una misma asignatura.	1, Si el alumno presenta acta de compromiso. 0, En caso contrario.

Factor	Variable	Valores de la variable
Académico	Tipo colegio de procedencia (TCP).	1, Si proviene de colegio particular. 0, En caso contrario.

Factor	Variable	Valores de la variable
Económico	Solicitud de recategorización (SR). Se refiere a la solicitud que realiza el alumno para que se le cambie de categoría de pago de pensiones (cambio por una de menor importe económico).	1, Si el alumno ha solicitado recategorización. 0, En caso contrario.
Económico	Ingreso familiar (IF). Es una variable que se ha optado por colocar en categorías, ya que así fue recabada al momento de la inscripción para el proceso de admisión.	Menos de S/. 2000. Entre S/. 2000 y S/. 4000. Más de S/. 4000.

FUENTE: ADAPTADO DE BARRENO Y ESPÍRITU , 2011

El artículo de (Berlanga Silvente, Figuera Gazo, & Pons Fanals, 2017) realiza un estudio, utilizando un modelo de regresión logística dirigido a buscar los factores determinantes de la permanencia del alumnado becado. El modelo obtenido muestra que el porcentaje de estudiantes correctamente clasificados es del 90,6%, por lo que el modelo clasifica adecuadamente a los estudiantes que persisten en sus estudios.

En el estudio se seleccionaron variables relativas al perfil personal y laboral, al perfil académico previo, motivacional y de rendimiento. De las 24 variables originalmente disponibles para el estudio solo 10 resultaron ser estadísticamente significativas y, por lo tanto, correlacionadas con la variable persistencia. Estas variables corresponden a: tasa de presentación (TP), tasa de rendimiento (TR), asignaturas aprobadas (AA), asignaturas presentadas (AP), nota media (NM), tasa de éxito (TE), nota de acceso (PAU), rama de conocimiento (RDC), asignaturas matriculadas (AM) y sexo (S) . Tiene sentido que en general, las variables académicas y de rendimiento se consideren factores predictivos de persistencia.

El estudio de (Gonzales Cam & Rodriguez Dominguez, 2017) presenta una propuesta de un Modelo de Business Intelligence para Identificar el perfil de deserción estudiantil en una universidad local, en el cual a través del procesamiento de diferentes fuentes de información y técnicas de minería de datos (árboles de decisión), les ha permitido identificar patrones que pueden ser analizados en modelos predictivos.

El estudio tomó como variables, consideradas como factores de riesgo, a las clasificadas como Académicas, Demográficas y las familiares, las mismas que identifican el grado de correlación, tal como se muestra en la tabla 04.

TABLA 04: VARIABLES CONSIDERADAS COMO FACTORES DE RIESGO

Variable	Factor de Riesgo	Clasificación	Correlación
ACAD01	Carrera	Académico	Alta
ACAD02	Semestre Cursado	Académico	Alta
ACAD03	Materias cursadas	Académico	Media
ACAD04	Modalidad de Ingreso	Académico	Alta
ACAD05	Materias perdidas	Académico	Alta
ACAD06	Horario o Programación	Académico	Media
ACAD07	Promedio	Académico	Alta
DEMD01	Edad	Demográfico	Media
DEM02	Ciudad de procedencia	Demográfico	Baja
DEM03	Domicilio	Demográfico	Media
DEM04	Sexo	Demográfico	Baja
DEM05	Estado Civil	Demográfico	Baja
FAM01	Nivel de estudios del Padre	Familiar	Baja
FAM02	Nivel de estudios de la Madre	Familiar	Baja
FAM03	Ingresos	Familiar	Alta
FAM04	Vive con sus padres	Familiar	Baja

FUENTE: GONZALES Y RODRÍGUEZ , 2017

En el trabajo de (Gonzales Cam & Rodriguez Dominguez, 2017) concluyeron que, a través de las herramientas de Minería de Datos, se puede crear modelos predictivos que ayuden al descenso de la tasa de deserción de 15% a 10%, además podría incrementar la recaudación en S/. 1,492,000 soles. El estudio de Gonzales y Rodríguez permitió evaluar el Valor Neto Actual el cual pasaría de 20,604,899 a 35,961,391 de soles, teniendo un TIR de 4005%, haciendo rentable el proyecto con una alta recuperación de la inversión.

4.2 Selección de la metodología

El presente trabajo de investigación usará la metodología CRISP-DM, según (Salcedo, Galeano, & Rodriguez, 2010), es una de las más utilizadas actualmente, basada en estándares internacionales que le permiten contar con fases estructuradas y confiables para los usuarios.

4.3 Aplicación de la metodología

4.3.1 Comprensión del negocio

En base a la metodología se propondrán los objetivos del negocio, una evaluación de la situación actual, luego se definirá los objetivos de la minería de datos y finalmente el plan de proyecto.

4.3.1.1 Objetivos del negocio

La institución educativa tiene como misión formar profesionales íntegros, impulsando caminos de superación.

Mantiene una estrategia de calidad e innovación académica, con acompañamiento a los estudiantes hasta su graduación.

Tiene como objetivo reducir la tasa de deserción.

Indicadores:

- Total Enrollment (students) – Total de estudiantes
- Ratio de retención (% alumnos)

4.3.1.2 Evaluación de la situación actual

a. Personal

La institución cuenta con área de desarrollo de sistema que se encarga de dar soporte y mantenimiento a los sistemas actuales, además desarrollan reportes a demanda, los especialistas conocen las estructuras de datos y la lógica del negocio. No existe un área que mantenga o verifique la calidad de datos, mucho menos especialistas que puedan desarrollar modelos de analítica haciendo uso de minería de datos, por lo cual se deberá considerar si algunas de las posiciones a considerar en el proyecto deberán mantenerse.

b. Datos

La institución cuenta con datos sistematizados del proceso académico, además registros de la información financiera, los cuales serán tomados en el análisis. También se identifican otros sistemas con información que puede ampliar el alcance del proyecto en el futuro y encontrar otros factores que puedan incidir en el problema de deserción, estos sistemas corresponden a los de CRM, AULA VIRTUAL, ENCUESTAS ACADEMICAS, REDES SOCIALES, entre otros.

c. Riesgo

El principal riesgo es no contar con el apoyo de alta dirección, lo cual puede ocasionar falta de compromiso de las áreas involucradas en el proyecto. Otros factores de riesgo están asociados los factores económicos y de tiempo. Desde el

punto de vista técnico, otro aspecto fundamental, es que los datos presenten escasa calidad o cobertura, que no permitan un análisis adecuado para predecir la deserción de los alumnos.

4.3.1.3 Definir Objetivos de la minería de datos

Generar un modelo predictivo que permita correlacionar las variables que afectan la permanencia estudiantil en el IES CIBERTEC Sede Norte, con información basada en datos recopilados en la matrícula del 2019-1.

Selección de un modelo predictivo que pueda ser implementado por la institución.

Identificar probabilidad que un alumno decida desertar.

Criterios de éxito:

- Identificar variables más representativas
- Seleccionar un modelo predictivo que supere un nivel de precisión del 80%

4.3.1.4 Plan de proyecto

En la tabla 05 se muestra las fases de la metodología CRISP-DM, la cual nos presenta un plan para la elaboración del proyecto de minería de datos

TABLA 05: PLAN DE PROYECTO DE ANALÍTICA DE DATOS

Fase	Tiempo	Recursos	Riesgos
Comprensión del negocio	5 semanas	Todos los analistas	Cambio económico
Comprensión de los datos	16 semanas	Todos los analistas	Problemas de datos, problemas tecnológicos
Preparación de los datos	12 semanas	Asesor de minería de datos, tiempo de análisis de base de datos	Problemas de datos, problemas tecnológicos
Modelado	8 semanas	Asesor de minería de datos, tiempo de análisis de base de datos	Problemas de tecnología, incapacidad para encontrar un modelo adecuado
Evaluación	4 semanas	Todos los analistas	Cambio económico, incapacidad para implementar resultados
Distribución	4 semanas	Asesor de minería de datos, tiempo de análisis de base de datos	Cambio económico, incapacidad para implementar resultados

FUENTE: ELABORACIÓN ADAPTADA DEL CRISP-MD IBM , 2012

4.3.2 Comprensión de los datos

4.3.2.1 Recolección inicial

Para la recolección de datos se ha considerado los siguientes sistemas:

- Sistema académico
- Sistema Financiero

Los siguientes formularios hacen referencia a los registros que formaran parte del análisis del presente estudio.

FIGURA 38: FORMULARIO DE ACTUALIZACIÓN DE DATOS PERSONALES

Datos de identificación La actualización de estos datos se hará efectiva, previa verificación, a las 24 horas de su registro por este formulario.

Nombres *

Apellido paterno *

Apellido materno *

Documento *

Fecha nacimiento * - - Sexo

Estado civil

Datos de domicilio

Vía Dirección * Urbanización

Dirección *

Distrito *

Provincia Departamento País

País *

Teléfono casa * Celular

Email personal

Nacionalidad

País nacimiento

Ciudad nacimiento

FUENTE: SISTEMA INTRANET SÓCRATES CIBERTEC, 2019

FIGURA 39: FORMULARIO DE INFORMACIÓN DE DATOS DE ALUMNOS

Alumno: código alumno

Generales

Nombre	nombres y apellidos	foto
Fecha de nacimiento	01/01/1900	
Sexo	masculino	
Teléfono		
Dirección		
Distrito	Independencia	
Nacionalidad		
Casilla de e-mail		

Admisión

Modalidad de estudios	Carreras Técnicas
Estudios	Carrera
Modalidad de ingreso	Directo
Ciclo de ingreso	
Código de postulante	
Proceso	

Colegio de procedencia

Nombre	
Dirección	
Año de egreso	

FUENTE: SISTEMA INTRANET SÓCRATES CIBERTEC, 2019

FIGURA 40: FORMULARIO DE INFORMACIÓN DE NOTAS ACTUALES

Modalidad de estudio: Ciclo académico :

Notas actuales
Consulte sus notas en el presente ciclo académico.

Alumno :
Estudio :
Nivel : 06 Total créditos : Tipo matrícula :
Nº asignaturas : Est. matrícula :

NOTAS POR ASIGNATURA								
	Código	Descripción	Sección	Grupo	Veces	Nivel	Nota	Existen Notas
☐	00000	PSICOLOGÍA GENERAL	0001	01	1	06	00.0	Si
☐	00001	PSICOLOGÍA GENERAL	0001	02	1	06	00.0	Si
☐	00002	PSICOLOGÍA GENERAL	0001	03	1	06	00.0	Si
☐	00003	PSICOLOGÍA GENERAL	0001	04	1	06	00.0	Si
☐	00004	PSICOLOGÍA GENERAL	0001	05	1	06	00.0	Si

Observaciones

[Comparativo](#) [Ver horario](#) [Ver Todas](#) [Impresión](#)

FUENTE: SISTEMA INTRANET SÓCRATES CIBERTEC, 2019

FIGURA 41: FORMULARIO DE INFORMACIÓN DE ASISTENCIA

Modalidad de estudio: Ciclo académico :

Inasistencias de alumnos
Consulte sus inasistencias en el presente ciclo académico

Alumno :
Estudio :

ASIGNATURA			CLASES						EFECTIVAS	
Código	Descripción	Programadas	Máx. Faltas	Dictadas	Asistidas	Tardanzas	Inasistidas	%	Inasistidas	%
		4	1	4	4	0	0	0 %	0	0 %
		4	1	4	4	0	0	0 %	0	0 %
		8	3	8	7	0	1	12.5 %	1	12.5 %
		4	1	4	4	0	0	0 %	0	0 %
		8	3	8	8	0	0	0 %	0	0 %

CLASES NO ASISTIDAS

Curso :

Fechas:

FUENTE: SISTEMA INTRANET SÓCRATES CIBERTEC, 2019

FIGURA 42: FORMULARIO DE COMPROBANTES DE PAGO

Consulta de comprobantes electrónicos
Esta consulta corresponde solo a documentos emitidos electrónicamente

Alumno:

Año: Mes: Tipo:

Num. documento	Moneda	Importe	Recargo	Descuento	Impuesto	Importe Cancelado	Saldo	Fecha Emisión	Fecha Vencimiento	Mora	Estado	Cuota	Periodo	Ver Documento
											COBRADO	1		
											PENDIENTE	2		
											PENDIENTE	3		
											PENDIENTE	4		
											PENDIENTE	5		

FUENTE: SISTEMA INTRANET SÓCRATES CIBERTEC, 2019

En base a la información contenida en el sistema se toma la data correspondiente al periodo 2019-1, para carreras técnicas de la sede Norte.

4.3.2.2 Descripción de los datos

En la tabla 06 se muestran las variables disponibles para el presente estudio, tomadas de los sistemas indicados en el punto anterior.

TABLA 06: DESCRIPCIÓN DE VARIABLES DE ANÁLISIS

TIPO	VARIABLE	DESCRIPCIÓN
ACADÉMICA	COD_LINEA_NEGOCIO	Línea de negocio de estudio
ACADÉMICA	COD_MODAL_EST	Modalidad de estudios
ACADÉMICA	COD_PERIODO	Periodo de análisis
ACADÉMICA	TIPO_INGRESO	Tipo de ingreso
ACADÉMICA	DES_ESTADO_MAT	Estado de matrícula
ACADÉMICA	DES_TIPO_MAT	Tipo de matrícula
ACADÉMICA	FACULTAD	Facultad de estudio, corresponde a la escuela de estudio
ACADÉMICA	DES_PRODUCTO	Descripción de la carrera de estudio
ACADÉMICA	COD_CURRICULO	Código de currículo vigente
ACADÉMICA	NUM_CURSO	Cantidad de cursos matriculados
ACADÉMICA	APROBADO_REAL	Cantidad de cursos aprobados
ACADÉMICA	DESAPROBADO_REAL	Cantidad de cursos desaprobados
ACADÉMICA	PCT_DESAPROBADO	Porcentaje de cursos desaprobados
ACADÉMICA	NOTAS_CURSOS	Detalle de notas por curso
ACADÉMICA	AVG_NOTA_REAL	Promedio de notas total
ACADÉMICA	INASISTENCIAS_CURSOS	Detalle de Porcentaje de inasistencia por curso
ACADÉMICA	PCT_INASISTENCIA_REAL	Porcentaje de inasistencias total
ACADÉMICA	CICLO_ALUMNO	Ciclo de alumno
ACADÉMICA	TIPO_COLEGIO	Tipo de colegio de procedencia
ACADÉMICA	ANIO_EGRESO_COLEGIO	Año de egreso del colegio de procedencia
ACADÉMICA	SEMANA_RETIRO	Semana en la que se retiró el alumno
ACADÉMICA	RETIRADO	Flag de retiro
DEMOGRAFICA	NOMBRE_UBIGEO	Distrito de procedencia
DEMOGRAFICA	ZONA	Zona de procedencia (Agrupación de NOMBRE_UBIGEO)
DEMOGRAFICA	PROVINCIA	Provincia de procedencia
DEMOGRAFICA	SEXO_ALUMNO	Sexo del alumno
DEMOGRAFICA	EDAD_ALUMNO	Edad de alumno (Derivada de la fecha nac.)
ECONÓMICA	CANT_BOLETAS	Cantidad de boletas canceladas

FUENTE: ELABORACIÓN PROPIA

4.3.2.3 Exploración o análisis inicial de los datos

En base a la información obtenida en el periodo de análisis 2019-1, el cual contiene 5653 registros de la sede Norte, se procede a realizar la exploración preliminar de los datos.

a. Análisis univariable

- PCT_INASISTENCIA_REAL

Tal como se muestra en la figura 43, la variable tiene como media 19% para el porcentaje de inasistencia, con una desviación del 20,6%. No se toman en cuenta 4 registros.

FIGURA 43: HISTOGRAMA DE FRECUENCIA PARA LA VARIABLE PCT_INASISTENCIA_REAL

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- EDAD_ALUMNO

Tal como se muestra en la figura 44, la variable tiene como media 21,98 años para la edad de los alumnos, con una desviación de 4,4 años. Existen 19 instancias que representa el 0,3% que no ha sido considerado por falta de datos o valores anómalos.

FIGURA 44: HISTOGRAMA DE FRECUENCIA PARA LA VARIABLE EDAD_ALUMNO

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- AVG_NOTA_REAL

Tal como se muestra en la figura 45, la variable tiene como media 13,17 para la nota promedio de los alumnos, con una desviación de 5.

FIGURA 45: HISTOGRAMA DE FRECUENCIA PARA LA VARIABLE AVG_NOTA_REAL

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- **NUM_CURSO**

Tal como se muestra en la figura 46, la variable tiene como media 6,83 para el número de cursos que llevan los alumnos en el periodo de análisis, con una desviación de 1,28. No se toman en cuenta 4 registros.

FIGURA 46: HISTOGRAMA DE FRECUENCIA PARA LA VARIABLE NUM_CURSO

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- **APROBADO_REAL**

Tal como se muestra en la figura 47, la variable tiene como media 5,1 para el número de cursos aprobados, con una desviación de 2,6. No se toman en cuenta 4 registros.

FIGURA 47: HISTOGRAMA DE FRECUENCIA PARA LA VARIABLE APROBADO_REAL

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- **CICLO_ALUMNO**

Tal como se muestra en la figura 48, se observa que el mayor porcentaje de alumnos pertenecen al primer ciclo con un 40,01% del total y en último lugar se tiene al sexto ciclo con el 7,84%.

FIGURA 48: DISTRIBUCIÓN DE ALUMNOS POR CICLO

FUENTE: ELABORACIÓN PROPIA

- **DES_PRODUCTO**

Tal como se muestra en la figura 49, se observa que las carreras de “Computación e Informática”, “Administración de Empresas” y “Diseño Gráfico” son las que se ubican en las tres primeras posiciones.

FIGURA 49: DISTRIBUCIÓN DE ALUMNOS POR CARRERA

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- FACULTAD

Tal como se muestra en la figura 50, se observa que el mayor porcentaje de alumnos pertenecen a las escuelas de “Gestión y Negocios”, “Diseño” y “Tecnologías de la Información”, con un 36,92%, 24,80% y 23,32% respectivamente. En último lugar se tiene a la escuela de “Salud” con un 2,02%.

FIGURA 50: DISTRIBUCIÓN DE ALUMNOS POR FACULTAD

FUENTE: ELABORACIÓN PROPIA

- RETIRADO

Tal como se muestra en la figura 51, se observa que los que decidieron retirarse representa el 11,09%.

FIGURA 51: DISTRIBUCIÓN DE ALUMNOS POR CONDICIÓN DE RETIRO

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- NOMBRE_UBIGEO

Tal como se muestra en la figura 52, se observa que los principales distritos de procedencia de los alumnos corresponden a SMP, COMAS, LOS OLIVOS e INDEPENDENCIA, los cuales superan el 9% para cada caso.

FIGURA 52: DISTRIBUCIÓN DE ALUMNOS POR LUGAR DE PROCEDENCIA

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

- TIPO_INGRESO

Tal como se muestra en la figura 53, se observa que el tipo de ingreso PID (Programa de Ingreso Directo) representa el 96,85% del total de alumnos.

FIGURA 53: DISTRIBUCIÓN DE ALUMNOS POR TIPO DE INGRESO

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

b. Análisis Bivariable

En la figura 54 se muestra la relación entre el “porcentaje de inasistencia” con la variable dependiente “retirado”, se puede observar la concentración de registros de los que decidieron retirarse se encuentran los alumnos con porcentaje de inasistencia entre el 40% y el 80% y para los que se mantuvieron en la institución están entre el 5% al 20% de porcentaje de inasistencia.

FIGURA 54: DIAGRAMA DE CAJAS RETIROS POR PCT_INASISTENCIA_REAL

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

En la figura 55 se muestra la relación entre el “edad” con la variable dependiente “retirado”, se puede observar la concentración de registros se mantienen en las edades entre los 19 a los 25 años. Se muestra un valor atípico en el registro 3751 con un alumno de menos de 10 años, esto refleja que la calidad de esta variable no está garantizada. De igual forma se debería revisar los valores outlayers que se encuentran en la variable edad.

FIGURA 55: DIAGRAMA DE CAJAS RETIROS POR EDAD_ALUMNO

FIGURA 56: DIAGRAMA DE CAJAS RETIROS POR AVG_NOTA_REAL

FUENTE: SOFTWARE IBM SPSS – ELABORACIÓN PROPIA

c. Variables categóricas

FIGURA 57: DIAGRAMA DE BARRAS RETIROS POR SEXO_ALUMNO

FUENTE: ELABORACIÓN PROPIA

FIGURA 58: DIAGRAMA DE BARRAS RETIROS POR FACULTAD

FUENTE: ELABORACIÓN PROPIA

FIGURA 59: DIAGRAMA DE BARRAS RETIROS POR DES_PRODUCTO

FUENTE: ELABORACIÓN PROPIA

4.3.2.4 Verificación de la calidad de los datos

Identificación de valores anómalos, datos vacíos, datos incompletos, errores en el registro

TABLA 07: RESULTADO DE AUDITORÍA DE DATOS

Campo	Medida	Mín.	Máx.	Media	Desv. est.	Sesgo	Exclusivo	Válido
COD_LINEA_NEGOCIO	Nominal	--	--	--	--	--	1	5653
COD_MODAL_EST	Nominal	--	--	--	--	--	1	5653
COD_PERIODO	Nominal	--	--	--	--	--	1	5653
TIPO_INGRESO	Nominal	--	--	--	--	--	15	5653
DES_ESTADO_MAT	Nominal	--	--	--	--	--	4	5653
DES_TIPO_MAT	Nominal	--	--	--	--	--	3	5653
FACULTAD	Nominal	--	--	--	--	--	6	5653
DES_PRODUCTO	Nominal	--	--	--	--	--	26	5653
NOMBRE_UBIGEO	Nominal	--	--	--	--	--	78	5653
PROVINCIA	Nominal	--	--	--	--	--	26	5653
COD_CURRICULO	Nominal	1	27	--	--	--	20	5653
NUM_CURSO	Nominal	1	10	--	--	--	10	5649
NUM_APROB	Nominal	0	10	--	--	--	11	5649
APROBADO_REAL	Nominal	0	10	--	--	--	11	5649
DESAPROBADO_REAL	Nominal	0	9	--	--	--	10	5649
PCT_DESAPROBADO	Continuo	0	1.167	0.221	0.324	1.432	--	5649
AVG_NOTA_REAL	Continuo	0	19.904	13.177	5.054	-1.523	--	5653
PCT_INASISTENCIA_REAL	Continuo	0	1	0.19	0.206	1.887	--	5649
CANT_BOLETAS	Nominal	0	12	--	--	--	13	5653
CICLO_ALUMNO	Nominal	1	6	--	--	--	6	5653
SEXO_ALUMNO	Nominal	--	--	--	--	--	2	5653
EDAD_ALUMNO	Nominal	8	58	--	--	--	36	5634
TIPO_COLEGIO	Nominal	--	--	--	--	--	7	5644
ANIO_EGRESO_COLEGIO	Nominal	0	2019	--	--	--	44	5569
SEMANA_RETIRO	Nominal	-1	15	--	--	--	10	5653
RETIRADO	Nominal	--	--	--	--	--	2	5653

FUENTE: ELABORACIÓN PROPIA

TABLA 08: RESULTADO DE CALIDAD DE DATOS

Campo	Valores atípicos	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio blanco	en	Valor vacío
COD_LINEA_NEGOCIO	--	100	5653	0	0	0	0	0
COD_MODAL_EST	--	100	5653	0	0	0	0	0
COD_PERIODO	--	100	5653	0	0	0	0	0
TIPO_INGRESO	--	100	5653	0	0	0	0	0
DES_ESTADO_MAT	--	100	5653	0	0	0	0	0
DES_TIPO_MAT	--	100	5653	0	0	0	0	0
FACULTAD	--	100	5653	0	0	0	0	0
DES_PRODUCTO	--	100	5653	0	0	0	0	0
NOMBRE_UBIGEO	--	100	5653	0	0	0	0	0

ZONA	--	100	5653	0	0	0	0
PROVINCIA	--	100	5653	0	0	0	0
COD_CURRICULO	--	100	5653	0	0	0	0
NUM_CURSO	--	99.929	5649	4	0	0	0
NUM_APROB	--	99.929	5649	4	0	0	0
APROBADO_REAL	--	99.929	5649	4	0	0	0
DESAPROBADO_REAL	--	99.929	5649	4	0	0	0
PCT_DESAPROBADO	0	99.929	5649	4	0	0	0
AVG_NOTA_REAL	0	100	5653	0	0	0	0
PCT_INASISTENCIA_REAL	155	99.929	5649	4	0	0	0
CANT_BOLETAS	--	100	5653	0	0	0	0
CICLO_ALUMNO	--	100	5653	0	0	0	0
SEXO_ALUMNO	--	100	5653	0	0	0	0
EDAD_ALUMNO	--	99.664	5634	19	0	0	0
TIPO_COLEGIO	--	99.841	5644	0	9	9	0
ANIO_EGRESO_COLEGIO	--	98.514	5569	84	0	0	0
SEMANA_RETIRO	--	100	5653	0	0	0	0
RETIRADO	--	100	5653	0	0	0	0

FUENTE: ELABORACIÓN PROPIA

4.3.3 Preparación de los datos

4.3.3.1 Datos relevantes

A través de nodo de selección de características de SPSS Modeler, se procede a configurar la importancia de las variables, se establece como punto de corte marginal 0.85:

FIGURA 60: CONFIGURACIÓN DE SELECCIÓN DE CARACTERÍSTICAS

FUENTE: ELABORACIÓN PROPIA

En la figura 61 se muestra las variables seleccionadas por el software, basado en la relación existe con la variable dependiente “Retiro”. Se excluyen las variables, SEXO_ALUMNO, EDAD_ALUMNO, TIPO_COLEGIO y ANIO_EGRESO_COLEGIO.

FIGURA 61: RESULTADO DE SELECTOR DE CARACTERÍSTICAS

Clasificación	Campo	Medida	Importancia	Valor
☒	1 DESAPROBADO...	Continuo	Importante	1,0
☒	2 PCT_DESAPROBA...	Continuo	Importante	1,0
☒	3 AVG_NOTA_REAL	Continuo	Importante	1,0
☒	4 PCT_INASISTENC...	Continuo	Importante	1,0
☒	5 APROBADO_REAL	Continuo	Importante	1,0
☒	6 CICLO_ALUMNO	Continuo	Importante	1,0
☒	7 CANT_BOLETAS	Continuo	Importante	1,0
☒	8 NUM_CURSO	Continuo	Importante	1,0
☒	9 DES_TIPO_MAT	Nominal	Importante	1,0
☒	10 DES_PRODUCTO1	Nominal	Importante	1,0
☒	11 DES_PRODUCTO	Nominal	Importante	1,0
☒	12 FACULTAD	Nominal	Importante	0,999
☒	13 ZONA	Nominal	Importante	0,852
☐	14 SEXO_ALUMNO	Nominal	Sin import...	0,379
☐	15 EDAD_ALUMNO	Continuo	Sin import...	0,248
☐	16 TIPO_COLEGIO	Nominal	Sin import...	0,171

Campos seleccionados: 13 Total de campos disponibles: 17

☒ > 0,85 ☒ <= 0,85 ☐ < 0,84

1 Campos representados

Campo	Medida	Motivo
☐ ANIO_EG...	Continuo	Coefficiente de variación por debajo de umbral

FUENTE: ELABORACIÓN PROPIA

4.3.3.2 Depuración de datos no relevantes

Se identifican los datos a excluir del análisis y la razón de la misma

TABLA 09: DATOS NO RELEVANTES

Variable	Motivo
COD_LINEA_NEGOCIO	Categoría única demasiado grande
COD_MODAL_EST	Categoría única demasiado grande
COD_PERIODO	Categoría única demasiado grande
TIPO_INGRESO	Categoría única demasiado grande
PROVINCIA	Categoría única demasiado grande
SEMANA_RETIRO	Demasiados valores perdidos
SEXO_ALUMNO	Baja relación con variable dependiente
EDAD_ALUMNO	Baja relación con variable dependiente
TIPO_COLEGIO	Baja relación con variable dependiente
ANIO_EGRESO_COLEGIO	Baja relación con variable dependiente

FUENTE: ELABORACIÓN PROPIA

4.3.3.3 Estructuración de datos

Se procede a estandarizar los datos para que puedan ingresar al análisis

a. AGRUPACIÓN

Para el análisis se realizaron las siguientes agrupaciones de datos

- **AVG_NOTA_REAL**

En la cual se calcula el promedio de notas de todos los cursos que tiene el alumno.

TABLA 10: NOTAS POR CURSO

PROMEDIO DE NOTAS	
Curso 1	Nota
Curso 2	Nota
Curso 3	Nota
Curso 4	Nota
Curso 5	Nota
Curso 6	Nota

FUENTE: ELABORACIÓN PROPIA

- **PCT_INASISTENCIA_REAL**

En el cual se hace el cálculo del porcentaje de inasistencia de todas las sesiones de clase para los diferentes cursos del alumno.

TABLA 11: INASISTENCIAS POR CURSO

PORCENTAJE INASISTENCIA TOTAL	
Curso 1	$\#_inas/\#_sesiones * 100$
Curso 2	$\#_inas/\#_sesiones * 100$
Curso 3	$\#_inas/\#_sesiones * 100$
Curso 4	$\#_inas/\#_sesiones * 100$
Curso 5	$\#_inas/\#_sesiones * 100$
Curso 6	$\#_inas/\#_sesiones * 100$

FUENTE: ELABORACIÓN PROPIA

- **ZONA**

Se procede a agrupar los distritos de procedencia de los alumnos, de acuerdo a la proximidad con la sede

TABLA 12: AGRUPACIÓN DE LUGAR DE PROCEDENCIA POR ZONA

ZONA_1				
Comas	Los Olivos			
Independencia	San Martín de Porres			
ZONA_2				
Ancón	Puente Piedra	Cercado de Lima		
Carabayllo	Ventanilla			
ZONA_3				
Asia	Cercado del Callao	Huaral	Lurigancho	Rímac
Ate Vitarte	Chaclacayo	Imperial	Lurín	San Borja
Aucallama	Chancay	Jesús María	Magdalena del Mar	San Isidro
Barranco	Chorrillos	La Molina	Mala	San Juan de Lurigancho
Bellavista	Cieneguilla	La Perla	Miraflores	San Juan de Miraflores
Breña	El Agustino	La Victoria	Pachacamac	San Luis
Carmen de la Legua Reynoso	Huacho	Lince	Pueblo Libre	San Miguel
Villa El Salvador	Santa Anita	Santa Rosa	Santiago de Surco	Surquillo
Villa María del Triunfo	Santa Cruz de Flores	Santa Rosa de Quives	Sayán	Otros

FUENTE: ELABORACIÓN PROPIA

b. NORMALIZACIÓN

Se procede a definir etiquetas a los valores de las variables FACULTAD y DES_PRODUCTO

TABLA 13: ETIQUETAS DE LA VARIABLE FACULTAD

FACULTAD	Etiqueta
ESCUELA DE DISEÑO	DI
ESCUELA DE TECNOLOGIAS DE LA INFORMACION	TI
ESCUELA DE GESTION Y NEGOCIOS	GN
ESCUELA DE INGENIERIA	IN
ESCUELA DE SALUD	SA
ESCUELA DE COMUNICACIONES	CO

FUENTE: ELABORACIÓN PROPIA

TABLA 14: ETIQUETAS DE LA VARIABLE PRODUCTOS (CARRERAS)

DES_PRODUCTO (Carrera)	Etiqueta
ADMINISTRACIÓN DE BANCA Y FINANZAS SN	1
ADMINISTRACIÓN DE EMPRESAS SN	2
ADMINISTRACIÓN DE NEGOCIOS BANCARIOS Y FINANCIEROS SN	3
ADMINISTRACIÓN DE NEGOCIOS INTERNACIONALES SN	4
ADMINISTRACION DE RECURSOS HUMANOS SN	5
ADMINISTRACIÓN DE REDES Y COMUNICACIONES SN	6
ADMINISTRACIÓN Y SISTEMAS SN	7
ANIMACIÓN DIGITAL SN	8
COMPUTACIÓN E INFORMÁTICA SN	9
COMUNICACIÓN Y ARTE DIGITAL SN	10
CONTABILIDAD SN	11
DIBUJO DE ARQUITECTURA Y OBRAS CIVILES SN	12
DISEÑO DE INTERIORES SN	13
DISEÑO DE VIDEOJUEGOS SN	14
DISEÑO GRÁFICO SN	15
ELECTRICIDAD INDUSTRIAL SN	16
GESTIÓN DE LA CONSTRUCCION CIVIL SN	17
GESTIÓN DE LOGISTICA SN	18
GESTION DE RECURSOS HUMANOS SN	19
MARKETING SN	20
MECANICA AUTOMOTRIZ Y AUTOTRONICA SN	21
MECATRÓNICA AUTOMOTRIZ SN	22
MECATRÓNICA INDUSTRIAL SN	23
REDES Y COMUNICACIONES SN	24
SEGURIDAD Y PREVENCIÓN DE RIESGOS SN	25
TRADUCCIÓN E INTERPRETACIÓN SN	26

FUENTE: ELABORACIÓN PROPIA

A continuación, se procede a estandarizar las variables categóricas nominales y ordinales convirtiéndolas en variables dummy dicotómicas (valores 0, 1). Esto permitirá realizar el análisis con el algoritmo de regresión logística.

En la tabla 15 se muestra la codificación dummy de las variables categóricas.

TABLA 15: CODIFICACIÓN DE VARIABLES CATEGÓRICAS

Codificaciones de variables categóricas							
		Frecuencia	Codificación de parámetro				
			(1)	(2)	(3)	(4)	(5)
FACULTAD	CO	24	1,000	,000	,000	,000	,000
	DI	158	,000	1,000	,000	,000	,000
	GN	223	,000	,000	1,000	,000	,000
	IN	64	,000	,000	,000	1,000	,000
	SA	11	,000	,000	,000	,000	1,000
	TI	139	,000	,000	,000	,000	,000
ZONA	ZONA_1	386	1,000	,000			
	ZONA_2	150	,000	1,000			
	ZONA_3	83	,000	,000			

FUENTE: ELABORACIÓN PROPIA

4.3.3.4 Integración de datos estructurados y no estructurados

No corresponde, debido a que no se cuentan con datos no estructurados

4.3.3.5 Formateo de datos

Para resolver el desbalanceo de datos de la variable objetivo, se define una directiva de equilibrado de datos:

FIGURA 62: CONFIGURACIÓN DE DIRECTIVA DE EQUILIBRADO DE DATOS

Configuración		Anotaciones
Directivas de equilibrado:		
Factor	Condición	
0.1324833425179099	RETIRADO = "NO"	
1.0	RETIRADO = "SI"	

FUENTE: ELABORACIÓN PROPIA

Antes de equilibrar

FIGURA 63: RECUENTO DE RETIROS ANTES DE EQUILBRAR DATOS

FUENTE: ELABORACIÓN PROPIA

Después de equilibrar

FIGURA 64: RECUENTO DE RETIROS DESPUÉS DE EQUILBRAR LOS DATOS

FUENTE: ELABORACIÓN PROPIA

4.3.4 Modelado

4.3.4.1 Selección de modelos propuestos

La propuesta plantea el uso de modelos de clasificación. En base a los estudios realizados se muestra que el uso de los algoritmos de regresión logística, árboles de decisión, etc. son los aplicables a la predictibilidad de la deserción.

Como primera aproximación, se procede a evaluar los diferentes modelos que presenta el nodo de clasificación automática de SPSS Modeler. La figura 65 muestra los resultados obtenidos, en el cual se identifica que los modelos CHAID y REGRESION LOGISTICA tienen niveles de precisión por encima del 96,3%.

FIGURA 65: RESULTADO DEL CLASIFICADOR AUTOMÁTICO

	Modelo	Tiempo de generación	Precisión general (%)	Nº de campos utilizados	Área debajo de la curva
CHAID 1	< 1	96,602	4	0,988
REGRESIÓN LOGÍSTICA 1	< 1	96,359	10	0,987
DISCRIMINANTE 1	< 1	96,359	8	0,997
QUEST 1	< 1	96,117	8	0,962
ÁRBOL C&R 1	< 1	96,117	9	0,959
RED NEURONAL 1	< 1	96,117	10	0,987

FUENTE: ELABORACIÓN PROPIA

4.3.4.2 Elaboración de plan de pruebas

Con apoyo de la herramienta se establecen tres particiones de datos, en el cual el 50% se asignará a Entrenamiento, el 30% para comprobación y el 20% para validación del modelo.

FIGURA 66: PARTICIONAMIENTO DE DATOS

The screenshot shows the 'Configuración' (Configuration) tab in the SPSS Modeler interface. It displays the following settings for data partitioning:

- Campo de partición:** Partición
- Particiones:**
 - ☐ Entrenamiento y comprobación
 - ☒ Entrenamiento, comprobación y validación
- Tamaño de partición de entrenamiento:** 50. Etiqueta: Entrenamiento. Valor = "1_Entrenamiento"
- Tamaño de partición de comprobación:** 30. Etiqueta: Comprobación. Valor = "2_Comprobación"
- Tamaño de partición de validación:** 20. Etiqueta: Validación. Valor = "3_Validación"
- Tamaño total:** 100%

FUENTE: ELABORACIÓN PROPIA

4.3.4.3 Construcción del modelo

a. Herramienta:

Como herramientas se ha seleccionado SPSS Modeler y SPSS Statistic, sin embargo, el presente estudio no sesga el uso de otras herramientas.

A continuación, se procede a evaluar los modelos de REGRESION LOGISTICA CHAID, para ello se diseña la ruta "Modelo_desercion" en SPSS Modeler.

FIGURA 67: RUTA “MODELO_DESERCION” EN SPSS

FUENTE: ELABORACIÓN PROPIA

4.3.4.3.1 Regresión Logística

De acuerdo a IBM ⁹ la regresión logística es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal, pero toma un campo objetivo categórico en lugar de uno numérico. En el caso del estudio propuesto, la variable objetivo corresponde a la categoría RETIRO: “SI” o “NO”.

⁹ consultado de IBM disponible en https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/modeling_nodes.html

FIGURA 68: CONFIGURACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

Análisis	
Análisis de matricula2.sav (13/10/2019 07:06:25 PM)	
Número de registros: 633	
Precisión del análisis: 96,524%	
Campos	
Destino	RETIRADO
Entradas	FACULTAD
	ZONA
	NUM_CURSO
	APROBADO_REAL
	DESAPROBADO_REAL
	PCT_DESAPROBADO
	AVG_NOTA_REAL
	PCT_INASISTENCIA_REAL
	CANT_BOLETAS
	CICLO_ALUMNO
Configuración de creación	
Utilizar los datos en particiones: true	
Partición: Partición	
Calcular importancia de predictor: true	
Calcular puntuaciones de propensión en bruto: falso	
Procedimiento: Binomial	
Tipo de modelo: Efectos principales	
Incluir constante en ecuación: true	
Modo: Simple	
Método binomial: Introducir	
Entradas categóricas:	
Resumen de entrenamiento	
Algoritmo: Regresión logística	
Tipo de modelo: Clasificación	
Aplicación: IBM® SPSS® Modeler 18	
Tiempo transcurrido para la generación del modelo: 0 horas, 0 minutos, 3 segundos	

FUENTE: ELABORACIÓN PROPIA

El análisis de Regresión Logística muestra en la figura 69 las variables de mayor importancia de predicción.

FIGURA 69: IMPORTANCIA DEL PREDICTOR EN MODELO DE REGRESIÓN LOGÍSTICA.

FUENTE: ELABORACIÓN PROPIA

Como se puede observar en la tabla de clasificación el pronóstico para los alumnos que se retiran acertó en un 96,2%. (Se toma 610 instancias de la fuente de datos).

TABLA 16: TABLA DE CLASIFICACIÓN

Tabla de clasificación					
	Observado		Pronosticado		
			RETIRADO		Porcentaje correcto
			NO	SI	
Paso 1	RETIRADO	NO	290	7	97,6
		SI	12	301	96,2
	Porcentaje global				96,9

FUENTE: ELABORACIÓN PROPIA

Para la validación del modelo obtenido se partitionaron los datos en tres grupos Entrenamiento, Comprobación y Validación, obteniendo en este último caso un porcentaje de acierto del 96.17%. Las métricas de evaluación son aceptables, para la evaluación del modelo se obtuvo para AUC 0,997 y para Gini 0,993

FIGURA 70: RESULTADOS DE PRECISIÓN Y MÉTRICAS EN MODELO DE REGRESIÓN LOGÍSTICA

Resultados para el campo de resultado RETIRADO						
Modelos individuales						
Comparando \$L-RETIRADO con RETIRADO						
'Partición'	1_Entrenamiento		2_Comprobación		3_Validación	
Correctos	611	96,52%	394	97,52%	251	96,17%
Erróneos	22	3,48%	10	2,48%	10	3,83%
Total	633		404		261	
Métricas de evaluación						
'Partición'	1_Entrenamiento		2_Comprobación		3_Validación	
Modelo	AUC	Gini	AUC	Gini	AUC	Gini
\$L-RETIRADO	0,994	0,988	0,996	0,992	0,997	0,993

FUENTE: ELABORACIÓN PROPIA

La tabla 17 muestra variables que intervienen en la fórmula de Regresión Logística

TABLA 17: VARIABLES DE LA ECUACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

		Variables en la ecuación					
		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1 ^a	FACULTAD			8,569	5	,128	
	FACULTAD(1)	-1,725	1,870	,851	1	,356	,178
	FACULTAD(2)	-,627	,979	,410	1	,522	,534
	FACULTAD(3)	,605	,802	,568	1	,451	1,831
	FACULTAD(4)	1,142	1,135	1,013	1	,314	3,133
	FACULTAD(5)	4,013	2,008	3,996	1	,046	55,334
	ZONA			,506	2	,776	
	ZONA(1)	,369	,947	,152	1	,697	1,447
	ZONA(2)	-,104	1,090	,009	1	,924	,901
	NUM_CURSO	-19,849	2559,799	,000	1	,994	,000
	APROBADO_REAL	19,971	2559,799	,000	1	,994	471311517,000
	DESAPROBADO_REAL	20,802	2559,799	,000	1	,994	1082083167,000
	PCT_DESAPROBADO	,586	7,106	,007	1	,934	1,796
	AVG_NOTA_REAL	-,158	,203	,606	1	,436	,854
	PCT_INASISTENCIA_REAL	10,742	3,027	12,595	1	,000	46273,319
	CANT_BOLETAS	-1,303	,309	17,762	1	,000	,272
	CICLO_ALUMNO	-,289	,230	1,583	1	,208	,749
	Constante	-3,487	5,373	,421	1	,516	,031
a. Variables especificadas en el paso 1: FACULTAD, ZONA, NUM_CURSO, APROBADO_REAL, DESAPROBADO_REAL, PCT_DESAPROBADO, AVG_NOTA_REAL, PCT_INASISTENCIA_REAL, CANT_BOLETAS, CICLO_ALUMNO.							

FUENTE: ELABORACIÓN PROPIA

El siguiente cuadro muestra un extracto de la predicción con el modelo de Regresión Logística y el porcentaje de precisión.

TABLA 18: EXTRACTO DE PREDICCIÓN CON EL MODELO DE REGRESIÓN LOGÍSTICA

FACULTAD	ZONA	NUM_CUR	APROBADO	DESAPROBADO	PCT_DESAPROBADO	AVG_NOTA_REAL	PCT_INASISTENCIA_REAL	CANT_BOLETAS	CICLO_ALUMNO	RETIRADO	Partición	SL-RETIRADO	SLP-RETIRADO	SLP-NO	SLP-SI
DI	ZONA_3	7,00	5,00	1,00	0,14	13,31	0,13	7,00	4,00	NO	3_Validació	NO	1,00	1,00	0,00
GN	ZONA_2	7,00	7,00	-	-	16,14	0,13	5,00	6,00	NO	3_Validació	NO	1,00	1,00	0,00
DI	ZONA_1	9,00	9,00	-	-	17,24	0,05	-	5,00	NO	3_Validació	NO	1,00	1,00	0,00
IN	ZONA_1	8,00	1,00	7,00	0,88	5,55	0,33	-	1,00	SI	3_Validació	SI	1,00	0,00	1,00
GN	ZONA_1	6,00	-	6,00	1,00	0,67	0,82	-	2,00	SI	3_Validació	SI	1,00	0,00	1,00
IN	ZONA_1	6,00	-	6,00	1,00	-	0,72	-	2,00	SI	3_Validació	SI	1,00	0,00	1,00
GN	ZONA_1	5,00	4,00	1,00	0,20	15,55	0,03	8,00	5,00	NO	3_Validació	NO	1,00	1,00	0,00
GN	ZONA_1	8,00	8,00	-	-	15,97	0,05	-	6,00	NO	3_Validació	NO	0,99	0,99	0,01
IN	ZONA_3	7,00	7,00	-	-	15,09	0,10	-	5,00	NO	3_Validació	NO	0,99	0,99	0,01
DI	ZONA_1	4,00	4,00	-	-	14,63	0,20	-	2,00	NO	3_Validació	NO	0,98	0,98	0,02
GN	ZONA_2	6,00	6,00	-	-	10,09	0,03	6,00	5,00	NO	3_Validació	NO	1,00	1,00	0,00
GN	ZONA_1	7,00	7,00	-	-	16,26	0,02	-	3,00	NO	3_Validació	NO	0,99	0,99	0,01
IN	ZONA_1	7,00	-	7,00	1,00	-	0,62	-	4,00	SI	3_Validació	SI	1,00	0,00	1,00
DI	ZONA_1	8,00	-	8,00	1,00	3,04	0,70	-	2,00	SI	3_Validació	SI	1,00	0,00	1,00
DI	ZONA_1	8,00	8,00	-	-	15,66	0,10	-	4,00	NO	3_Validació	NO	1,00	1,00	0,00
GN	ZONA_1	7,00	7,00	-	-	15,27	0,10	2,00	4,00	NO	3_Validació	NO	1,00	1,00	0,00
TI	ZONA_1	8,00	7,00	1,00	0,13	14,64	0,02	-	5,00	NO	3_Validació	NO	0,99	0,99	0,01
IN	ZONA_1	4,00	2,00	2,00	0,50	8,50	0,03	-	3,00	SI	3_Validació	NO	0,80	0,80	0,20
IN	ZONA_2	7,00	7,00	-	-	15,97	0,09	-	4,00	NO	3_Validació	NO	0,99	0,99	0,01
SA	ZONA_1	8,00	8,00	-	-	15,77	0,15	7,00	5,00	NO	3_Validació	NO	1,00	1,00	0,00
GN	ZONA_1	1,00	1,00	-	-	16,25	0,21	5,00	6,00	NO	3_Validació	NO	1,00	1,00	0,00
GN	ZONA_1	1,00	1,00	-	-	13,70	0,14	5,00	6,00	NO	3_Validació	NO	1,00	1,00	0,00
TI	ZONA_1	7,00	-	7,00	1,00	-	0,39	-	5,00	SI	3_Validació	SI	1,00	0,00	1,00
GN	ZONA_1	7,00	6,00	1,00	0,14	15,18	0,14	-	6,00	NO	3_Validació	NO	0,97	0,97	0,03
TI	ZONA_1	8,00	-	8,00	1,00	-	0,45	-	4,00	SI	3_Validació	SI	1,00	0,00	1,00
TI	ZONA_1	7,00	2,00	5,00	0,71	8,57	0,40	-	3,00	SI	3_Validació	SI	0,99	0,01	0,99
TI	ZONA_1	7,00	7,00	-	-	16,09	0,07	-	3,00	NO	3_Validació	NO	0,99	0,99	0,01
GN	ZONA_1	6,00	-	6,00	1,00	1,59	0,80	-	3,00	SI	3_Validació	SI	1,00	0,00	1,00

FUENTE: ELABORACIÓN PROPIA

4.3.4.3.2 Árbol de decisión CHAID

De acuerdo a IBM¹⁰, el nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas.

Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

A continuación, se especifica la configuración del modelo de Clasificación con el árbol de decisión CHAID.

FIGURA 71: CONFIGURACIÓN DEL ÁRBOL DE DECISIÓN CHAID

FUENTE: ELABORACIÓN PROPIA

¹⁰ Consultado de IBM disponible en

https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/modeling_nodes.html

Los resultados obtenidos con este modelo muestran que la variable predictor de mayor importancia es PCT_DESAPROBADO, seguida por CANT_BOLETAS, AVG_NOTA_REAL y CICLO_ALUMNO, tal como se muestra en la figura 72.

FUENTE: ELABORACIÓN PROPIA

Como se puede observar en la tabla de clasificación el pronóstico para los alumnos que se retiran acertó en un 94,1% (Se toma 620 instancias de la fuente de datos).

TABLA 19: TABLA DE CLASIFICACIÓN

Tabla de clasificación					
	Observado		Pronosticado		
			RETIRADO		Porcentaje correcto
			NO	SI	
Paso 1	RETIRADO	NO	321	12	96.4%
		SI	17	270	94.1%
	Porcentaje global				95.3%

FUENTE: ELABORACIÓN PROPIA

Para la validación del modelo obtenido se particionaron los datos en tres grupos: Entrenamiento, Comprobación y Validación, obteniendo en este último caso un porcentaje de acierto del 95,08%. Las métricas de evaluación son aceptables, para la evaluación del modelo se obtuvo para AUC 0,992 y para Gini 0,985.

FIGURA 73: RESULTADOS DE PRECISIÓN Y MÉTRICAS EN MODELO CHAID

Resultados para el campo de resultado RETIRADO							
Modelos individuales							
Comparando \$R-RETIRADO con RETIRADO							
'Partición'	1_Entrenamiento		2_Comprobación		3_Validación		
Correctos	608	94,7%	385	94,13%	251	95,08%	
Erróneos	34	5,3%	24	5,87%	13	4,92%	
Total	642		409		264		
Métricas de evaluación							
'Partición'	1_Entrenamiento		2_Comprobación		3_Validación		
Modelo	AUC	Gini	AUC	Gini	AUC	Gini	
\$R-RETIRADO	0,99	0,98	0,987	0,974	0,992	0,985	

FUENTE: ELABORACIÓN PROPIA

A continuación, se muestra las reglas generadas por el árbol CHAID, donde la predicción del retiro de alumnos se establece con un porcentaje considerable, se da en el nodo 16 (N16), nodo 10 (N10) y nodo 11 (N11). El nodo 10 es el que presenta mejor precisión, con un 100% de acierto para 220 instancias que decidieron desertar.

FIGURA 74: RESULTADOS DE PRECISIÓN Y MÉTRICAS EN MODELO CHAID

(N1) PCT_DESAPROBADO <= 0 [Moda: NO] => NO (191; 1,0)
 (N2) PCT_DESAPROBADO > 0 and PCT_DESAPROBADO <= 0,143 [Moda: NO] => NO (51; 0,941)
 (N3) PCT_DESAPROBADO > 0,143 and PCT_DESAPROBADO <= 0,400 [Moda: NO] (57)
 (N6) CANT_BOLETAS <= 0 [Moda: NO] (37)
 (N12) AVG_NOTA_REAL <= 13,769 [Moda: NO] => NO (23; 0,522)
 (N13) AVG_NOTA_REAL > 13,769 [Moda: NO] => NO (14; 0,929)
 (N7) CANT_BOLETAS > 0 [Moda: NO] => NO (20; 1,0)
 (N4) PCT_DESAPROBADO > 0,400 and PCT_DESAPROBADO <= 0,833 [Moda: SI] (68)
 (N8) CANT_BOLETAS <= 3 [Moda: SI] (57)
 (N14) AVG_NOTA_REAL <= 10,714 [Moda: SI] (42)
 (N16) CICLO_ALUMNO <= 2 [Moda: SI] => SI (34; 0,941)
 (N17) CICLO_ALUMNO > 2 [Moda: NO] => NO (8; 0,5)
 (N15) AVG_NOTA_REAL > 10,714 [Moda: NO] => NO (15; 0,6)
 (N9) CANT_BOLETAS > 3 [Moda: NO] => NO (11; 1,0)
 (N5) PCT_DESAPROBADO > 0,833 [Moda: SI] (248)
 (N10) CANT_BOLETAS <= 0 [Moda: SI] => SI (220; 1,0)
 (N11) CANT_BOLETAS > 0 [Moda: SI] => SI (28; 0,893)

FUENTE: ELABORACIÓN PROPIA

En la figura 75 se presenta el árbol CHAID generado

FIGURA 75: ÁRBOL GENERADO CON EL MODELO CHAID

FUENTE: ELABORACIÓN PROPIA

En el siguiente cuadro se muestra un extracto los resultados de la predicción con el árbol de clasificación CHAID, para la partición de validación.

TABLA 20: EXTRACTO DE PREDICCIÓN CON EL MODELO CHAID

FACULTAD	ZONA	NUM_CURS	APROBADO	DESAPROBA	PCT_DESAPR	AVG_NOTA	PCT_INASIST	CANT_BOLET	CICLO_ALUM	RETIRADO	Partición	\$R-RETIRAD	\$RC-RETIRAD
DI	ZONA_1	8.00	7.00	1.00	0.13	13.64	0.19	6.00	4.00	NO	3_Validación	NO	1
SA	ZONA_2	1.00	1.00	-	-	19.08	0.14	3.00	6.00	NO	3_Validación	NO	1
GN	ZONA_1	7.00	7.00	-	-	15.40	0.20	-	5.00	NO	3_Validación	NO	1
DI	ZONA_2	9.00	7.00	2.00	0.22	14.62	0.18	9.00	4.00	NO	3_Validación	NO	1
DI	ZONA_1	9.00	9.00	-	-	16.82	0.13	-	5.00	NO	3_Validación	NO	1
GN	ZONA_1	7.00	7.00	-	-	18.63	0.10	-	5.00	NO	3_Validación	NO	1
TI	ZONA_2	7.00	7.00	-	-	16.11	0.04	-	3.00	NO	3_Validación	NO	1
GN	ZONA_1	5.00	5.00	-	-	14.77	0.11	3.00	5.00	NO	3_Validación	NO	1
TI	ZONA_1	3.00	3.00	-	-	13.92	0.09	-	6.00	NO	3_Validación	NO	1
GN	ZONA_3	7.00	7.00	-	-	15.59	0.09	-	5.00	NO	3_Validación	NO	1
GN	ZONA_2	7.00	7.00	-	-	15.94	0.08	-	4.00	NO	3_Validación	NO	1
GN	ZONA_3	7.00	7.00	-	-	17.49	0.04	3.00	5.00	NO	3_Validación	NO	1
TI	ZONA_3	8.00	8.00	-	-	18.29	0.05	-	6.00	NO	3_Validación	NO	1
DI	ZONA_1	8.00	5.00	3.00	0.38	14.34	0.12	1.00	4.00	NO	3_Validación	NO	1
DI	ZONA_1	8.00	-	8.00	1.00	3.04	0.70	-	2.00	SI	3_Validación	SI	1
GN	ZONA_1	4.00	4.00	-	-	14.49	0.07	3.00	3.00	NO	3_Validación	NO	1
TI	ZONA_3	7.00	-	7.00	1.00	-	0.61	-	5.00	SI	3_Validación	SI	1
TI	ZONA_1	8.00	5.00	3.00	0.38	12.52	0.09	-	5.00	NO	3_Validación	NO	1
IN	ZONA_1	4.00	2.00	2.00	0.50	8.50	0.03	-	3.00	SI	3_Validación	NO	1
GN	ZONA_1	7.00	7.00	-	-	17.07	0.16	-	5.00	NO	3_Validación	NO	1
GN	ZONA_1	7.00	7.00	-	-	14.90	0.11	-	4.00	NO	3_Validación	NO	1
DI	ZONA_1	5.00	-	5.00	1.00	-	0.60	-	2.00	SI	3_Validación	SI	1
SA	ZONA_2	6.00	6.00	-	-	16.71	0.05	6.00	4.00	NO	3_Validación	NO	1
DI	ZONA_1	6.00	6.00	-	-	16.18	0.19	-	4.00	NO	3_Validación	NO	1
TI	ZONA_1	8.00	7.00	1.00	0.13	13.44	0.18	8.00	4.00	NO	3_Validación	NO	1
DI	ZONA_1	6.00	6.00	-	-	15.27	0.13	-	2.00	NO	3_Validación	NO	1
GN	ZONA_2	7.00	7.00	-	-	15.35	0.04	-	4.00	NO	3_Validación	NO	1

FUENTE: ELABORACIÓN PROPIA

4.3.5 Evaluación

A continuación, se realiza el comparativo de los modelos de regresión logística y árbol de decisión CHAID. En primer lugar, se realiza la matriz de confusión con los datos de evaluación y se tienen los resultados en las tablas. Como se puede observar la clasificación con el modelo de regresión logística tiene una mejor precisión con un 96.5% a comparación del árbol CHAID que obtiene una precisión de solo el 93.7%.

TABLA 21: TABLA DE CLASIFICACIÓN REGRESIÓN LOGÍSTICA (VALIDACIÓN)

Tabla de clasificación Regresión Logística					
	Observado		Pronosticado		
			RETIRADO		Porcentaje correcto
			NO	SI	
Paso 1	RETIRADO	NO	127	5	96.2%
		SI	4	123	96.9%
	Porcentaje global				96.5%

FUENTE: ELABORACIÓN PROPIA

TABLA 22: TABLA DE CLASIFICACIÓN ÁRBOL CHAID (VALIDACIÓN)

Tabla de clasificación Árbol CHAID					
	Observado		Pronosticado		
			RETIRADO		Porcentaje correcto
			NO	SI	
Paso 1	RETIRADO	NO	120	3	97.6%
		SI	13	117	90.0%
	Porcentaje global				93.7%

FUENTE: ELABORACIÓN PROPIA

Comparando los modelos se muestra una ligera inclinación hacia el modelo de Regresión Logística el cual tiene un AUC de 0,997 a comparación del árbol de clasificación CHAID con 0.991.

FIGURA 76: COMPARACIÓN DE RESULTADOS REGRESIÓN LOGÍSTICA Y ÁRBOL CHAID

FUENTE: ELABORACIÓN PROPIA

Con esta comparación podemos seleccionar el modelo de regresión logística como el que mejor para este estudio.

Implementación del modelo seleccionado

Basados en los factores encontrados en el modelo de regresión logística se elabora la fórmula que permita determinar el porcentaje de probabilidad de deserción, en la tabla 23 se muestran los coeficientes para cada variable considerada en el modelo.

TABLA 23: COEFICIENTES DE LA FORMULA GENERADA CON EL MODELO DE REGRESIÓN LOGÍSTICA

Fórmula F(y)
FACULTAD (1) * -1.7249556422578
FACULTAD (2) * -0.627201506910208
FACULTAD (3) * 0.604590173267217
FACULTAD (4) * 1.14202966621855
FACULTAD (5) * 4.01339462927418
ZONA (1) * 0.369443450482925
ZONA (2) * -0.104081534026688
NUM_CURSO * -19.8490970645744
APROBADO_REAL * 19.9710298281889
DESAPROBADO_REAL * 20.8021538783139
PCT_DESAPROBADO * 0.585515539555357
AVG_NOTA_REAL * -0.158178997398907
PCT_INASISTENCIA_REAL * 10.7423208096318
CANT_BOLETAS * -1.30290300011178
CICLO_ALUMNO * -0.288808747964289
Constante * -3.48669950683796

FUENTE: ELABORACIÓN PROPIA

El modelo logístico se constituye en relación entre la probabilidad de que ocurra el suceso, debido a que el agente presenta los valores (X x1, X x2 , , X x3):

FIGURA 77: FÓRMULA DE REGRESIÓN LOGÍSTICA

$$P[Y = 1/x_1, x_2, \dots, x_k] = \frac{1}{1 + e^{(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_k x_k)}}$$

FUENTE: BARRENO VEREAU, 2012

Para nuestro estudio se establece la fórmula del modelo de la siguiente manera

FIGURA 78: FÓRMULA DE REGRESIÓN LOGÍSTICA PARA LA PROBABILIDAD DE DESERCIÓN

$$P(\text{deserción}) = \frac{1}{1 + e^{-y}}$$

FUENTE: ADAPTADO DE BERLANGA SILVENTE, FIGUERA GAZO & PONS FANALS, 2017

A través de una implementación básica realizada en un documento Excel, se elabora

FIGURA 79: FÓRMULA IMPLEMENTADA EN EXCEL

Variables en la ecuación
		100%
FACULTAD	GN	
ZONA	ZONA_1	
NUM_CURSO	7.00	
APROBADO_REAL	3.00	
DESAPROBADO_REAL	4.00	
PCT_DESAPROBADO	0.57	
AVG_NOTA_REAL	7.71	
PCT_INASISTENCIA_REAL	0.53	
CANT_BOLETAS	0.00	
CICLO_ALUMNO	3.00	
Constante	1.00	
Σ		5.607

FUENTE: ELABORACIÓN PROPIA

Para la evaluación final se verifica los criterios de éxito del objetivo de la minería de datos:

- ✓ Se logra identificar las variables representativas para el modelo
- ✓ El modelo predictivo seleccionado, regresión logística, a logrado un nivel de precisión superior al 80%

5 EVALUACIÓN ECONÓMICA

Se procederá con el detalle de los montos que se necesitarán para llevar a cabo la implementación del modelo predictivo utilizando analítica de datos.

Se realizará el análisis financiero de la inversión y el impacto monetario en la organización con la implementación del modelo.

5.1 Análisis de la Situación Actual

Tal como se mencionó en capítulo 3.2, la sede norte es la que contribuye con la mayor cantidad de alumnos en el IES CIBERTEC y a su vez representa el mayor porcentaje de retiros para la institución. En base a la información del periodo 2019 I la sede norte alcanzó un nivel de deserción del 11.09%, lo cual corresponde a 627 alumnos que iniciaron sus clases y decidieron retirarse dentro del periodo académico. El ingreso promedio anual por alumno es de S/ 7,308.00

TABLA 24: SITUACIÓN ACTUAL DE LA SEDE NORTE

Estimación en base a información 2019 I-Sede Norte		
Total Matricula (# alumnos matriculados)	5653	
Total no retenidos 2019 I	627	11.09%
Ingreso promedio anual por alumno	S/ 7,308.00	

FUENTE: ELABORACIÓN PROPIA

5.2 Presupuesto del Proyecto

5.2.1 Consideraciones para la elaboración del presupuesto

Se establecerá el presupuesto considerando la contratación de un equipo de especialistas para la implementación de analítica de datos en la institución. Para la etapa de implementación, en el primer año se considera la contratación de un administrador de base de datos y un jefe de proyecto. Adicionalmente, se considera la adquisición de un servidor, equipos PCs para el desarrollo de actividades y las licencias correspondientes.

No se considerará los costos de alquileres ni de materiales, dado que están dentro del presupuesto administrativo propio de la institución.

Tal como se muestra en la figura 80, la infraestructura propuesta está compuesta por 3 servidores virtuales, alojados en el servidor físico adquirido, con sistema operativo Windows Server y tecnología de virtualización Hyper V. Los softwares a usar: SQL Server (Data Base, SSIS Integración de Servicios, SSRS Reporting Service), servicio de IIS para la publicación de web services y Power BI para la publicación de cuadros de mando.

FIGURA 80: INFRAESTRUCTURA PROPUESTA

FUENTE: ELABORACIÓN PROPIA

Cibertec al ser una institución educativa puede obtener costos en licenciamiento en las soluciones Microsoft, basado en el Campus Agreement, logrando descuentos de casi 80% de su precio de lista. Los costos de software Microsoft considerados en el presupuesto se han tomado considerando este tipo de licenciamiento.

Se establece un costo para aplicar planes de retención, que permitan retener a los estudiantes con alto grado de deserción, identificados con el modelo predictivo.

A continuación, se describe el presupuesto de implementación, considerando los recursos descritos anteriormente:

TABLA 25: PRESUPUESTO DE IMPLEMENTACIÓN

Hardware	Medida	Cantidad	Precio Unitario	Total (S/.)
Servidor Proliant DL380 Gen 10 (Procesador 2XGen 10 4114 Xeon-S FIO, RAM 2X32GB 2Rx4, DISCO 8X600GB SAS 10K SFF)	UN	1	36,108.00	36,108.00
PCs	UN	5	2,800.00	14,000.00
Multifuncional	UN	1	2,500.00	2,500.00
<i>Subtotal</i>				<i>52,608.00</i>

Software				
Suscripción Herramienta DM (Complemento Herramienta Free)	ANUAL	1	9,628.80	9,628.80
Licencia Hyper - V(Incluido en WinSer DataCenter)	UN	3	0.00	0.00
SQL Server (SQLSvrEntCore ALNG LicSAPk OLV 2Lic E 1Y Acdmc AP CoreLic)	ANUAL	3	5,035.06	15,105.18
Windows Server (WinSvrDCCore ALNG LicSAPk OLV 2Lic E 1Y Acdmc AP CoreLic)	ANUAL	3	401.20	1,203.60
PwrBIProOpenFac ShrdSvr SNGL SubsVL OLP NL Annual Acdmc Qlfd	ANUAL	30	124.37	3,731.16
<i>Subtotal</i>				<i>29,668.74</i>

Recursos especializados				
Administrador de BD (Solo 1er año)	UN	12	6,000.00	72,000.00
Analista Jr. Minería de datos	UN	12	3,500.00	42,000.00
Analista Sr. Minería de datos	UN	12	7,000.00	84,000.00
Jefe de Proyecto (Solo 1er año)	UN	12	8,000.00	96,000.00
<i>Subtotal</i>				<i>294,000.00</i>
Mobiliario				
Escritorios	UN			0
Sillas	UN			0
<i>Subtotal</i>				<i>0</i>
Gastos Administrativos-otros				
Plan de retención	UN	12	17,000.00	204,000.00
Local	UN			0
Servicios	UN			0
utiles	UN			0
<i>Subtotal</i>				<i>204,000.00</i>
Total Presupuesto				580,276.74

FUENTE: ELABORACIÓN PROPIA

5.2.2 Flujo de caja del Proyecto

De acuerdo al artículo de D2L¹¹ (Campos, 2017) se menciona como caso de éxito de Oral Roberts University (Oklahoma, Estados Unidos), en el cual con el uso de datos analíticos y estrategias de retención se llegó a aumentar entre el 15.5% y el 25.5% la tasa de retención.

En la entrevista a Francisco Rojas¹² (Gestión, 2017) se menciona que las nuevas tecnologías de almacenamiento, procesamiento y modelamiento matemático de datos permiten identificar a tiempo los móviles de deserción de los estudiantes para corregirlos. Ello lograría

¹¹ D2L es la empresa creadora de Brightspace, una plataforma online de enseñanza y aprendizaje fácil, flexible e inteligente

¹² Francisco Rojas CEO de Penta Analytics, empresa dedicada a brindar soluciones analíticas.

disminuir hasta en 13% la tasa de abandono de alumnos de primer año y hasta en 7% la de los alumnos más antiguos.

Basados en estos indicadores, el presente estudio evalúa escenarios para poder establecer el flujo de caja del proyecto y verificar la viabilidad del mismo:

TABLA 26: ESCENARIOS PROPUESTOS PARA EL ANÁLISIS ECONÓMICO

	Pesimista	Moderado	Optimista
↓ % Deserción	10.0%	9.8%	8.9%
% Tasa de Retención	10.0%	12.0%	20.0%
Retención con la herramienta de analítica	63	75	125
Recaudado x alumnos retenidos (anual)	458,211.60	549,853.92	916,423.20

FUENTE: ELABORACIÓN PROPIA

Otras consideraciones:

- Se establece que la implementación de analítica de datos para CIBERTEC demore 1 año. Para la evaluación de la factibilidad económica del proyecto se realizará una proyección a 5 años.
- No se tiene planeado realizar un préstamo para la implementación del presente proyecto, debido a que se está presupuestando como parte de las partidas correspondientes a proyectos de Cibertec.
- Se visualiza que los flujos de ganancia mantendrán una mejor rentabilidad.
- Para la comparación se utilizarán los escenarios pesimista y moderado.

5.2.2.1 Flujo de caja Escenario Pesimista

A continuación, en la tabla 27, se muestra el estado de resultado para el escenario PESIMISTA, en el cual se espera retener un 10% de probables desertores (63 alumnos retenidos al año), con un ahorro esperado de S/ 458,211.60. Los montos están expresados en soles (S/.):

TABLA 27: ESTADO DE RESULTADOS ESCENARIO PESIMISTA

	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Ahorro esperado		458,211.60	458,211.60	458,211.60	458,211.60	458,211.60
(-) costo software		29,668.74	29,668.74	29,668.74	29,668.74	29,668.74
Ahorro Utilidad Bruta		428,542.86	428,542.86	428,542.86	428,542.86	428,542.86
(-) Gastos administrativos		294,000.00	126,000.00	126,000.00	126,000.00	126,000.00
--- (-) Gastos para programas de Retención		204,000.00	204,000.00	204,000.00	204,000.00	204,000.00
(-) Depreciación		13,152.00	13,152.00	13,152.00	13,152.00	-

Ahorro (Utilidad) antes de interés e impuestos (EBIT)		(82,609.14)	85,390.86	85,390.86	85,390.86	98,542.86
Gastos Financieros		-	-	-	-	-
Ahorro (Utilidad) imponible		(82,609.14)	85,390.86	85,390.86	85,390.86	98,542.86
(-) Impuesto a las ganancias (30%)		-	25,617.26	25,617.26	25,617.26	29,562.86
Ahorro Neto (Utilidad neta)		(82,609.14)	59,773.60	59,773.60	59,773.60	68,980.00

Costos

540,820.74	372,820.74	372,820.74	372,820.74	359,668.74
------------	------------	------------	------------	------------

Flujo de caja libre

	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Ahorro (Utilidad) antes de interés e impuestos (EBIT)	-	(82,609.14)	85,390.86	85,390.86	85,390.86	98,542.86
Impuesto a las ganancias (30%)	-	-	25,617.26	25,617.26	25,617.26	29,562.86
EBIT-Impuesto=NOPAT	-	(82,609.14)	59,773.60	59,773.60	59,773.60	68,980.00
(+) Depreciación y amortización	-	13,152.00	13,152.00	13,152.00	13,152.00	-
(-) Cambio en capital de trabajo	-	-	-	-	-	-
(+) Valor Residual	-	-	-	-	-	-
Inversión en activos	(52,608.00)	-	-	-	-	-
FLUJO DE CAJA LIBRE	(52,608.00)	(69,457.14)	72,925.60	72,925.60	72,925.60	68,980.00

FUENTE: ELABORACIÓN PROPIA

De la tabla 27 podemos indicar los siguientes resultados de la evaluación económica con un escenario pesimista:

TABLA 28: EVALUACIÓN ECONÓMICA ESCENARIO PESIMISTA

Evaluación económica	
TMAR (Tasa Mínima Aceptable de Rendimiento)	10%
VAN	S/91,948.64
TIR	36%
PRI O PAYBACK DESCONTADO	3.01
PRI O PAYBACK	2.67
B/C	1.18
ROI	175%

FUENTE: ELABORACIÓN PROPIA

Teniendo como base la tasa mínima aceptable de rendimiento (TMAR) del 10%, los resultados obtenidos son favorables, lo cual confirma la viabilidad financiera del proyecto. De acuerdo al Payback descontado(actualizado) el retorno de la inversión se da después del

3er año de implementado. Se obtiene un TIR de 36% , ganando un 118% por cada sol invertido.

5.2.2.2 Flujo de caja Escenario Moderado

A continuación, en la tabla 29, se muestra el estado de resultado para el escenario MODERADO, en el cual se espera retener un 12% de probables desertores (75 alumnos retenidos al año), con un ahorro esperado de S/ 549,853.92. Los montos están expresados en soles (S/.):

TABLA 29: ESTADO DE RESULTADOS ESCENARIO MODERADO

	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Ahorro esperado		549,853.92	549,853.92	549,853.92	549,853.92	549,853.92
(-) costo software		29,668.74	29,668.74	29,668.74	29,668.74	29,668.74
Ahorro Utilidad Bruta		520,185.18	520,185.18	520,185.18	520,185.18	520,185.18
(-) Gastos administrativos		294,000.00	126,000.00	126,000.00	126,000.00	126,000.00
--- (-) Gastos de programas de Retención		204,000.00	204,000.00	204,000.00	204,000.00	204,000.00
(-) Depreciación		13,152.00	13,152.00	13,152.00	13,152.00	-
Ahorro (Utilidad) antes de interés e impuestos (EBIT)		9,033.18	177,033.18	177,033.18	177,033.18	190,185.18
Gastos Financieros		-	-	-	-	-
Ahorro (Utilidad) imponible		9,033.18	177,033.18	177,033.18	177,033.18	190,185.18
(-) Impuesto a las ganancias (30%)		2,709.95	53,109.95	53,109.95	53,109.95	57,055.55
Ahorro Neto (Utilidad neta)		6,323.23	123,923.23	123,923.23	123,923.23	133,129.63

Costos

540,820.74	372,820.74	372,820.74	372,820.74	359,668.74
------------	------------	------------	------------	------------

Flujo de caja libre

	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Ahorro (Utilidad) antes de interés e impuestos (EBIT)	-	9,033.18	177,033.18	177,033.18	177,033.18	190,185.18
(-) Impuesto a las ganancias (30%)	-	2,709.95	53,109.95	53,109.95	53,109.95	57,055.55
EBIT-Impuesto=NOPAT	-	6,323.23	123,923.23	123,923.23	123,923.23	133,129.63
(+) Depreciación y amortización	-	13,152.00	13,152.00	13,152.00	13,152.00	-
(-) Cambio en capital de trabajo	-	-	-	-	-	-
(+) Valor Residual	-	-	-	-	-	-
Inversión en activos	(52,608.00)	-	-	-	-	-
FLUJO DE CAJA LIBRE	(52,608.00)	19,475.23	137,075.23	137,075.23	137,075.23	133,129.63

FUENTE: ELABORACIÓN PROPIA

De la tabla 29 podemos indicar los siguientes resultados de la evaluación económica con un escenario moderado:

TABLA 30: EVALUACIÓN ECONÓMICA ESCENARIO MODERADO

Evaluación económica	
TMAR (Tasa Mínima Aceptable de Rendimiento)	10%
VAN	S/357,655.96
TIR	130%
PRI O PAYBACK DESCONTADO	1.31
PRI O PAYBACK	1.24
B/C	1.42
ROI	680%

FUENTE: ELABORACIÓN PROPIA

Teniendo como base la tasa mínima aceptable de rendimiento (TMAR) del 10%, los resultados obtenidos son favorables, lo cual confirma la viabilidad financiera del proyecto. De acuerdo al Payback descontado(actualizado) el retorno de la inversión se da después del 1er año de implementado, con un TIR del 130%, recuperando 142% por cada sol invertido.

En la comparación de ambos escenarios, se puede identificar que, aplicando la analítica y estrategias de retención, se obtiene como resultado una inversión viable, lo cual permitiría recuperar más de lo que actualmente se deja de percibir por los alumnos no retenidos.

Considerando que estos resultados no se darían si es que no se realiza una inversión para los planes deserción, se ha estipulado un presupuesto de S./204,000.00 los cuales involucra principalmente las siguientes áreas:

- Servicio Académico
- Actividades Sociales
- Estrategias de Apoyo

Cada una de estas áreas serán las que deben tener un plan de acuerdo con el perfil encontrado para el alumno o grupo de alumnos.

5.2.3 Resultados en Beneficio para Cibertec

Los beneficios que se pueden establecer para la presente propuesta son:

- ✓ Contar con la identificación temprana de los posibles desertores, que permita lograr eficiencia en la aplicación de los planes de retención, focalizando los recursos de manera adecuada.
- ✓ Mejora en la gestión asociada a los riesgos de deserción.
- ✓ Mejorar la rentabilidad de la institución, a través de los ingresos por la retención de alumnos con riesgo.
- ✓ La integración de la analítica de datos a los procesos del negocio, podría dar una ventaja competitiva en comparación a otras instituciones.
- ✓ Contar con una solución aplicada a otras realidades, como por ejemplo a otras sedes.

6 CONCLUSIONES

- El presente trabajo nos permitió identificar patrones con los cuales se pueden focalizar planes estratégicos de retención académicos, con una precisión mayor al 90% superando, el objetivo inicial colocado para la investigación.
- Contar con una metodología de analítica, como CRISP_DM, nos permitió orientar de manera correcta y ordenada el uso de los datos, teniendo en cuenta que existe un respaldo es diversos estudios que nos indica que han sido probadas por otras instituciones logrando un mejor performance al momento de tomar decisiones.
- En la evaluación económica, se ha encontrado que se tiene una cuota alta para la adquisición de hardware, lo cual podría ser reemplazado por servicios in cloud. En este caso se ha respetado el landscape de Cibertec, en cuanto al manejo de su infraestructura.
- Se debe considerar un presupuesto para los planes de retención, dado que estos tendrán diferentes objetivos y estarán focalizados a los perfiles que se puedan definir con los patrones que muestre la analítica de datos.
- Se ha podido esta que la herramienta por sí sola no es la solución inmediata, lo que realmente le da un valor real y preponderante es el análisis que se realiza a través a de los datos y la calidad de información que estos puedan mostrar, así como las diversas estrategias y acciones que ejecute la institución en base a los patrones que muestre la analítica.

Con lo mencionado no se busca forzar al estudiante ni a la institución de que continúen ese vínculo que busque cumplir el objetivo de ambos, sino por lo contrario, que la institución sea un ente que monitoree la evolución del mismo y sea una alternativa para poder identificar de manera oportuna aquel estudiante que presente una fisura en su correcto desarrollo profesional lo cual podría generar una alta probabilidad de deserción. Tal como se pudo observar en la data analizada y pruebas realizadas, las herramientas de analítica nos pueden proporcionar de manera simple un escenario para evaluar las diversas variables que influyen en el desarrollo de un estudiante mediante el modelo planteado en la solución, lo cual nos permite tomar medidas a tiempo y focalizar las acciones a tomar a fin de mitigar la alta probabilidad de ocurrir un evento puntual.

7 RECOMENDACIONES

El presente trabajo permitió evidenciar que existen patrones de información para identificar a posibles desertores. Lo cual ayuda a formular perfiles en los estudiantes y poder realizar un correcto seguimiento de los datos que puedan proporcionar el desarrollo del estudiante en el día a día y poder alertar a tiempo a la institución de los diversos factores que puedan alterar el desarrollo profesional y actuar para contribuir en la correcta continuidad del estudiante y cumplir con sus objetivos profesionales, sin embargo, debemos considerar lo siguiente:

- Se recomienda utilizar toda la información a la cual se tenga accesos en la institución, a fin de poder evaluar otras variables que puedan ser determinantes al momento de predecir si el alumno tiene o no una alta probabilidad de deserción. Se debe contar con un repositorio de datos que permita el almacenamiento de características sociales y culturales.
- Es importante contar un área y personal idóneo que analice la información, que se recaba en los diversos sistemas con los que se cuentan en la institución, así mismo continuar con diversas pruebas de analytics que permitan identificar otros perfiles que permitan a la institución poder aplicar diversas estrategias.
- La herramienta usada para analytics y predicción de resultados, SPSS, es solo una de las muchas que existen en el mercado, también se podría optar por buscar herramientas gratuitas y simples que le permitan tener una visión y panorama de los diversos resultados que se puedan obtener.
- Es primordial y necesario contar con data homogénea y limpia entre los diversos sistemas que se manejan en la institución, dado que durante las pruebas se encontró data incongruente, la cual de continuar así podría generar resultados engaños y a su vez perjudicar los modelos a obtener, para ello es importante contar con la metodología que servirá de guía para este fin.
- El estudio solo tomo la información de una sede por lo que el modelo propuesto en la solución debe ser evaluado, para que se pueda precisar y aplicar a otras sedes. Considerando que la inversión sería menor ya que se podría reutilizar los recursos presentados en la implementación.

- Teniendo en cuenta que el comportamiento de las variables analizadas puede variar, probablemente se deba ajustar el modelo.
- El estudio solo analizó la deserción que se presenta dentro del periodo académico iniciado, pero también existe una oportunidad de la aplicación de analítica de datos para los desertores que no inician un nuevo ciclo, ya que también se cuenta con datos históricos de estos estudiantes y puedan permitir construir un modelo en particular, incluyendo otras variables.
- El estudio tampoco incluyó a los desertores precoces, conocidos como “no show”, que son aquellas personas que, a pesar de haber realizado su registro como ingresantes, no iniciaron el 1er ciclo de estudios. Se recomienda se pueda evaluar los datos que se tenga de estos clientes, asociándolos con los del sistema CRM y otros y permita realizar un análisis predictivo para estos casos.
- Se recomienda la revisión de la infraestructura de sistemas que cuenta CIBERTEC con la finalidad de encontrar una mejora en como gestiona los servicios que tiene.
- Contar con una estructura y protocolo de seguridad para los datos sensibles, dado que para la aplicación de los planes se debe tomar con nombre y apellido al alumno con la finalidad de aplicar el plan de retención correcto.
- Al manejar información relevante y muy sensible CIBERTEC debe contar con un gobierno de datos eficiente para lograr integrar un concepto holístico del conocimiento de la información, Asimismo el control de la gestión de los datos con un enfoque general de las necesidades del negocio, manteniendo una coordinación interactiva entre las áreas estratégicas de la institución.

8 GLOSARIO DE TERMINOS

IES. Sigla que representa a los Institutos de Educación Superior del país, para nuestro estudio hacemos referencia específicamente a CIBERTEC.

MINEDU. Representa al Ministerio de Educación es el sector del Poder Ejecutivo encargado de la educación en la República del Perú.

EES. Sigla que representa a las Escuelas de Educación Superior

EEST. Sigla que representa a las Escuelas de Educación Superior Tecnológica

EESP - Sigla que representa a las Escuelas de Educación Superior Pedagógica

DRELM - Sigla que representa a la Dirección Regional de Educación de Lima Metropolitana

CBC - Sigla que representa a las Condiciones Básicas de Calidad con las que debe cumplir la Institución de Educación Superior

INEI. El Instituto Nacional de Estadística e Informática, es un organismo constitucional autónomo del Perú. Se encarga de dirigir los sistemas nacionales de estadística y también de informática del país. Es también encargado de los censos de población, de vivienda, de empresas, agrarios, universitarios, etc.

Algoritmos. Es un conjunto reescrito de instrucciones o reglas bien definidas, ordenadas y finitas que permite realizar una actividad mediante pasos sucesivos que no generen dudas a quien lo ejecute.

Tupla. Una tupla es una secuencia de valores agrupados. Una tupla sirve para agrupar, como si fueran un único valor, varios valores que, por su naturaleza, deben ir juntos. El tipo de datos que representa a las tuplas se llama tuple. El tipo tuple es inmutable: una tupla no puede ser modificada una vez que ha sido creada.

Overfitting. En aprendizaje automático, el sobreajuste (también es frecuente emplear el término en inglés overfitting) es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado

Hipótesis. es un enunciado no verificado, una vez refutado o confirmado dejará de ser hipótesis y sería un enunciado verificado. La hipótesis es una conjetura científica que requiere una contrastación con la experiencia.

OLAP. s el acrónimo en inglés de procesamiento analítico en línea. Es una solución utilizada en el campo de la llamada Inteligencia de negocios cuyo objetivo es agilizar la consulta de grandes cantidades de datos

Data warehouse. En el contexto de la informática, un almacén de datos es una colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza

DM. Data Mining o Minería de Datos, es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos

JDBC. Java Database Connectivity (en español: Conectividad a bases de datos de Java), más conocida por sus siglas JDBC, es una API que permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java, independientemente del sistema operativo donde se ejecute

GNU. Es un sistema operativo de tipo Unix, así como una gran colección de programas informáticos que componen al sistema, desarrollado por y para el Proyecto GNU y auspiciado por la Free Software Foundation.

Machine Learning. El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan

Framework. Un framework, entorno de trabajo o marco de trabajo es un conjunto estandarizado de conceptos, prácticas y criterios para enfocar un tipo de problemática particular que sirve como referencia, para enfrentar y resolver nuevos problemas de índole similar

Python. Es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional

Código Abierto. es un modelo de desarrollo de software basado en la colaboración abierta. Se enfoca más en los beneficios prácticos que en cuestiones éticas o de libertad que tanto se destacan en el software libre.

SINEACE. El Sistema Nacional de Evaluación, Acreditación y Certificación de la Calidad Educativa – Sineace es un organismo técnico especializado, adscrito al Ministerio de Educación, creado en el 2006 a través de la Ley N°28740. Su finalidad es garantizar a la sociedad que las instituciones educativas públicas y privadas ofrezcan un servicio de calidad, y que los trabajadores peruanos estén altamente calificados para la labor que realizan

PRONABEC. Programa Nacional de Becas y Crédito Educativo (PRONABEC), que promueve el acceso de jóvenes a distintos tipos de educación superior, captando y administrando becas donadas por instituciones de educación superior públicas y privadas.

CRM. La customer relationship management, más conocida por sus siglas CRM, puede tener varios significados: Administración basada en la relación con los clientes, un modelo de gestión de toda la organización, basada en la satisfacción del cliente

Focus Group. es una técnica cualitativa de estudio de las opiniones o actitudes de un público, utilizada en ciencias sociales y en estudios comerciales

TIR. La tasa interna de retorno o tasa interna de rentabilidad de una inversión es la media geométrica de los rendimientos futuros esperados de dicha inversión, y que implica por cierto el supuesto de una oportunidad para "reinvertir"

9 REFERENCIAS

- Apaza, E., & Huamán, F. (2012). *Factores determinantes que inciden en la deserción de los estudiantes universitarios*. Universidad Peruana Union. San Martin, Perú: Apuntes Universitarios. Revista de Investigación. Obtenido de <http://www.redalyc.org/articulo.oa?id=467646124005>
- Aquino, A., Molero-Castillo, G., & Rojano, R. (2015). Hacia un nuevo proceso de minería de datos centrado en el usuario. *Pistas Educativas*(114), 272-291.
- Barreno Vereau, E. V. (2012). *Análisis comparativo de modelos de clasificación en el estudio de la deserción universitaria*. Universidad de Lima, Lima, Perú. Obtenido de <http://repositorio.ulima.edu.pe/handle/ulima/2505>
- Berlanga Silvente, V., Figuera Gazo, M. P., & Pons Fanals, E. (2017). *MODELO PREDICTIVO DE PERSISTENCIA UNIVERSITARIA: ALUMNADO CON BECA SALARIO*. Madrid, España: Educación XX1. doi:<https://doi.org/10.5944/educxx1.20193>
- Bravo Castilo, M., & Mejía Giraldo, A. (2010). *Los retos de la educación superior en Colombia: una reflexión sobre el fenómeno de la deserción universitaria*. Universidad de San Buenaventura. Cali, Colombia: Revista Educación En Ingeniería. doi:<https://doi.org/10.26507/rei.v5n10.101>
- Campos, W. (14 de Febrero de 2017). *Datos analíticos para mejorar las tasas de retención universitaria*. Obtenido de D2L: <https://www.d2l.com/es/blog/datos-analiticos-para-mejorar-las-tasas-de-retencion-universitaria/>
- Canales, A., & De los Rios, D. (2007). *Factores explicativos de la deserción universitaria. Calidad en la Educación*. Universidad de Santiago de Chile. Santiago de Chile, Chile: Revista Calidad en la Educación. doi:<https://doi.org/10.31619/caledu.n26.239>
- Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2004). Deserción Estudiantil universitaria: Una aplicación de modelo de duración. *Lectura de Economía*.

- CIBERTEC. (2018). *Proyecto Educativo Institucional*. Obtenido de <https://www.cibertec.edu.pe/wp-content/uploads/2018/12/Proyecto-Educativo-Institucional-2017-2021v3.pdf>
- Continental, E. d. (15 de Mayo de 2019). *El Buho*. Obtenido de El Buho: <https://elbuho.pe/2019/05/desercion-universitaria-alcanza-el-30-en-peru-y-42-en-latinoamerica/>
- Decreto supremo N° 010-2017. (06 de febrero de 2019). Resolución Viceministerial. Obtenido de <http://www.minedu.gob.pe/ley-de-institutos/pdf/rvm-n-020-2019-aprueba-la-cbc-de-licenciamiento.pdf>
- Deepak, A. (2014). *Analytics based decision making*. Journal of Indian Business Research. Bangalore, India: Emerald Group Publishing Limited. doi:<https://doi.org/10.1108/JIBR-09-2014-0062>
- Del Carmen Parrino, M. (2014). *Factores intervinientes en el fenómeno de la deserción universitaria*. Revista Argentina de Educación Superior, ISSN-e 1852-8171, N°. 8. Obtenido de <https://dialnet.unirioja.es/servlet/articulo?codigo=4753784>
- Eckert, K. B., & Suénaga, R. (2015). *Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos*. Centro de Información Tecnológica, La Serena, Chile. Obtenido de <http://www.redalyc.org/articulo.oa?id=373544192002>
- Ferreira, M. M., Avitabile, C., Botero Álvarez, J. H., & Urzúa, S. (2017). *La educación superior en América*. Grupo Bnaco Mundial. Obtenido de <https://openknowledge.worldbank.org/bitstream/handle/10986/26489/211014ovSP.pdf?sequence=5&isAllowed=y>.
- Fontalvo, W., Castillo, M., & Polo, S. (2014). Análisis comparativo entre las características. Escenarios.
- Gestión. (11 de Julio de 2017). *El 27% de ingresantes a universidades privadas abandonan su carrera en primer año de estudios*. Obtenido de DIARIO GESTIÓN: <https://gestion.pe/tendencias/management-empleo/27-ingresantes-universidades-privadas-abandonan-carrera-primer-ano-estudios-139168-noticia/>

- Gonzales Cam, C., & Rodriguez Dominguez, C. (2017). *Propuesta de un Modelo de Business Intelligence para Identificar el perfil de deserción estudiantil en la Universidad Científica del Sur*. Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Perú. Obtenido de <http://hdl.handle.net/10757/622749>
- González, L. (2005). *Estudio sobre la repitencia*. Recuperado el 17 de Noviembre de 2019, de Digital Observatory for higher education in: <http://unesdoc.unesco.org/images/0014/001400/140087s.pdf>
- Grossman, R., Kamath, C., Kegelmeyer, P., Kumar, V., & Namburu, R. (2013). *Data Mining for Scientific and Engineering Applications*. Springer Science & Business Media. Obtenido de <https://books.google.com.pe/books?id=h1zmBwAAQBAJ>
- Haro Rivera, S., Zúñiga Lema, L., Vera Rojas, L., Meneses Freire, A., & Escudero Villa, A. (2018). *Métodos de clasificación en minería de datos meteorológicos*. Ecuador: Escuela Superior Politécnica de Chimborazo. Obtenido de <http://dspace.espoch.edu.ec/handle/123456789/9395>
- Hernandez, J., Ramirez, M. J., & Ferri, C. (2004). *Introducción a la minería de datos*. Madrid: Pearson Educación.
- IBM. (2011). *IBM Business Analytics*. Nueva York, U.S.A.: IBM Corporation. Obtenido de http://public.dhe.ibm.com/software/data/sw-library/cognos/services/pdfs/implementation/business_analytics_solutions_implementation_method_data_sheet.pdf
- IBM. (2012). *Manual CRISP-DM de IBM SPSS*. U.S.A.: IBM Corporation. Obtenido de <ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf>
- International Business Machines (IBM). (12 de Junio de 2015). *Metodología fundamental para la ciencia de datos*. Recuperado el 13 de Octubre de 2019, de <https://www.ibm.com/downloads/cas/6RZMKDN8>
- Jiménez Toledo, J. A., & Timarán Pereira, S. R. (2015). *Caracterización de la deserción estudiantil en educación superior con minería de datos*. Institución Universitaria CESMAG. Colombia: Revista Tecnológica ESPOL. Obtenido de <http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/453>

Ley N°30512. (2 de Noviembre de 2016). Ley de Insititutos y escuelas de educación superior y de la carrera publica de sus docentes.

Lopera Oquendo, C. (2008). *Determinantes de la deserción universitaria en la Facultad de Economía Universidad del Rosario*. Universidad del Rosario. Bogota, Colombia: Borradores de Investigación, N° 95. Obtenido de [https://repository.urosario.edu.co/bitstream/handle/10336/3740/Borrador_Economia_95\[1\].pdf?sequence=8](https://repository.urosario.edu.co/bitstream/handle/10336/3740/Borrador_Economia_95[1].pdf?sequence=8)

Moine, J. M. (2013). *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo*. La Plata, Argentina: Universidad Nacional de La Plata.

Moine, J. M. (2013). *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo (Tesis doctoral)*. Buenos Aires: Universidad Nacional de la Plata.

PRONABEC. (2014). *Plan de Gestión Institucional del PRONABEC MINEDU 2014-2018*.

PRONABEC. (2015). *Costos Académicos de Instituciones Educativas de Beca 18*. Lima: Programa Nacional de Beca y Crédito Educativo-Ministerio de Educación.

PRONABEC. (2016). *Plan de Gestión Institucional del Programa Nacional de Becas y Crédito Educativo 2016 – 2021 Rumbo al Bicentenario*.

Quintero Velasco, I. (2016). *Análisis de las causas de deserción universitaria*. Universidad Nacional Abierta y a Distancia UNAD, BOGOTA, COLOMBIA. Obtenido de <https://repository.unad.edu.co/bitstream/10596/6253/1/23783211.pdf>

Rivo, E., de la Fuente, J., Rivo, Á., Garcia, E., Cañizares, M., & Gil, P. (2012). Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clinical and Translational Oncology*(14), 73-79.

Rodríguez Lagunas, J., & Hernández Vázquez, J. M. (2008). *La deserción escolar universitaria en México. La experiencia de la Universidad Autónoma Metropolitana Campus Iztapalapa*. Costa Rica: Actualidades Investigativas en Educación. Obtenido de

<http://132.248.9.34/hevila/Actualidadesinvestigativaseneducacion/2008/vol8/no1/15.pdf>

- Salcedo, O., Galeano, R., & Rodriguez, L. (2010). Metodología crisp para la implementación Data Warehouse. *Revista Tecnura*, 14(26), 35-48. Obtenido de <https://revistas.udistrital.edu.co/index.php/Tecnura/article/view/6685>
- Saldaña Villa, M., & Barriga, O. A. (2010). *Adaptación del modelo de deserción universitaria de Tinto a la Universidad Católica de la Santísima Concepción, Chile*. *Revista de Ciencias Sociales*, 16(4), 616-628.
- Sebastián, D., Donoso, G., & Oscar, A. (2010). *Iniciativas de retención de estudiantes de educación superior*. Universidad de Talca. Chile: Revista Calidad en la Educación. doi:<https://doi.org/10.31619/caledu.n33.138>
- Sumathi, S., & Sivanandam, S. (2006). *Introduction to Data Mining and its Applications. Studies in Computational Intelligence*. Heidelberg: Springer-Verlag.
- Tinto, V. (1989). *Definir la deserción: una cuestión de perspectiva*. México: Revista de Educación Superior, 71.
- Valle Benavides, A. R. (2017). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones*. Universidad de Sevilla, Sevilla, España. Obtenido de <http://hdl.handle.net/11441/63201>
- Vélez, A., & López, D. (2009). *Estrategia para vencer la deserción universitaria*. Universidad de la Sabana. La Sabana, Colombia: Educación y Educadores. Obtenido de <https://educacionyeducadores.unisabana.edu.co/index.php/eye/article/view/555/648>