

**UNIVERSIDAD RICARDO PALMA
FACULTAD DE INGENIERÍA
PROGRAMA DE TITULACIÓN POR TESIS
ESCUELA PROFESIONAL DE INGENIERÍA
INFORMATICA**



**MODELO PREDICTIVO DE DESERCIÓN
UNIVERSITARIA DE LA CARRERA DE INGENIERÍA
INFORMATICA EN LA UNIVERSIDAD RICARDO
PALMA**

**TESIS
PARA OBTENER EL TÍTULO PROFESIONAL DE
INGENIERA INFORMATICA**

PRESENTADO POR:

**Bach. GALVEZ CHAMBILLA MELISSA BEATRIZ
Bach. FLORES CORNEJO KATHERINE BRIGGITE**

ASESOR: Lic. RENZO WARTHON VARELA

LIMA - PERÚ

AÑO: 2015

DEDICATORIA

Dedico esta tesis a mi familia, por su infinito amor e incondicional apoyo que se ha hecho presente en situaciones favorables tanto como en las difíciles, en especial a mi hijo Daniel que es mi mayor motivo de superación.

Melissa Galvez Chambilla

Dedico este trabajo a mis padres por su amor, trabajo y sacrificio en todos estos años, quienes me acompañaron durante todo este esfuerzo para arribar a la meta.

A mi hermano quien me alentó para continuar y lograr este objetivo.

Katherine Flores Cornejo

AGRADECIMIENTOS

Agradecemos a nuestros padres por la educación que nos brindaron, a los profesores por haber sido guías en este recorrido universitario y a todos los que nos dieron su apoyo en el transcurso de la elaboración de este proyecto.

RESUMEN

La deserción universitaria se ha convertido en un problema importante a ser investigado. La tasa de deserción ha llegado a constituir uno de los principales indicadores de eficiencia interna dentro de cualquier institución de educación. Investigar las causas de la deserción con metodologías adecuadas que permitan predecir esta, contribuye a la toma de decisiones dentro de la gestión de la unidad académica.

El objetivo del presente proyecto consiste en investigar y proponer una metodología que permita identificar en forma automática a los estudiantes con mayor riesgo de deserción de las carreras de Ingeniería Informática en la Universidad Ricardo Palma.

Para la implementación de este proyecto se adoptó la metodología CRISP-DM que estructura el proceso de minería de datos en seis fases, que interactúan entre ellas de forma iterativa.

Se evaluaron los modelos de Árboles de decisión y Clustering para analizar el comportamiento de los estudiantes, evaluando factores como el rendimiento del alumno, condición social y aspectos socioeconómicos. La exactitud de los modelos es calculada a partir de la información que brindó la Oficina Central de Informática y Cómputo de la Universidad Ricardo Palma, en la cual se realizó una transformación y simulación de algunas variables para mayor efectividad del modelo.

Palabras claves: Deserción universitaria, Minería de datos, Modelo predictivo, árbol de decisión.

ABSTRACT

The college desertion has become an important problem to be investigated. The desertion rate has come to be one of the main indicators of internal efficiency in any educational institution. The research of the causes of desertion with appropriate methodologies to predict, contributes to decision making in the academic management

The objective of this project is to investigate and propose a methodology to automatically identify students at high risk of dropping out of the Computer Engineering career at the Ricardo Palma University.

To implement this project, we used the CRISP-DM methodology to structure the data mining process into six phases, which interact with each other iteratively.

It was applied the models of decision trees and clustering to analyze the behavior of the students, evaluating factors such as student performance, social status and socio-economic aspects. The accuracy of the models is calculated from the information provided the Central Office of Information and Computation at the Ricardo Palma University, in which it was applied a transformation and simulation of some variables for greater effectiveness and performance of the model.

Keywords: Desertion University, Data Mining, Predictive Model, Decision tree

Índice Temático

RESUMEN	IV
ABSTRACT	V
INTRODUCCIÓN	1
Antecedentes del Problema	2
Definición del Problema	6
Propósito del Proyecto	7
Importancia del Proyecto	9
Justificación del Proyecto	9
Definición de Términos	10
Resumen	13
CAPITULO I: MARCO TEORICO	14
Introducción	14
Revisión de la literatura	14
Conclusiones	45
CAPITULO II: DESARROLLO DEL PROYECTO	46
1. Alcance del Proyecto	46
Alcance del producto	46
Criterios de aceptación del producto	47
Estructura de Desglose del Trabajo y Entregables (EDT)	48
Exclusiones del proyecto	51
Restricciones del proyecto	52
Supuestos del proyecto	53
2. Modelado de Negocio	54

3. Requerimientos del Producto/Software	59
3.1 Requerimientos funcionales	59
3.2 Matriz de trazabilidad	60
3.3 Requerimientos No Funcionales	61
4. Arquitectura del Producto/Software	62
1. Vista de Despliegue	62
2. Vista de implementación	63
4. Vista lógica	66
5. Vista de datos	67
6. Vista de casos de uso	69
5. Diseño de la solución	70
6. Evaluación de Resultados	103
CONCLUSIONES	113
RECOMENDACIONES	115
BIBLIOGRAFIA	116
ANEXO 1: Flujo alternativo de CUS Transformar y cargar datos	120
ANEXO 2: Informe de Variables	124
ANEXO 4: Estrategia de Ejecución	127
ANEXO 5: Informe de Colección de Datos	128
ANEXO 6: Informe de Calidad de Datos	129
ANEXO 7: Variable Socio Económica	132
ANEXO 8: Instrucciones de Limpieza de Datos	134
ANEXO 9: Variables usadas en Modelo Predictivo	135
ANEXO 10: Confirmación de Carga de Datos	136

Índice de Ilustraciones

Ilustración 1: Diagrama de Ishikawa Causas de la deserción universitaria	4
Ilustración 2: Alumnos desertores por periodo matriculado	6
Ilustración 3: Diagrama de Ubicación del proyecto	8
Ilustración 4: Encuesta realizada por la KDnuggets en el año 2007	26
Ilustración 5: Cuatro niveles de detalle de la metodología CRISP-DM	27
Ilustración 6: Categoría de las variables	30
Ilustración 7: El proceso CRISP DM	33
Ilustración 8: Mapa de la Minería de Datos [Bibliografía Nro 10]	40
Ilustración 9: Representación de una base de datos	42
Ilustración 10: Proceso ETL	44
Ilustración 11: EDT del proyecto	48
Ilustración 12: Diagrama de caso de uso del negocio, TO BE	54
Ilustración 13: Diagrama del proceso propuesto (To Be)	55
Ilustración 14: Matriz de trazabilidad	60
Ilustración 15 Diagrama de despliegue	63
Ilustración 16: Diagrama de implementación	64
Ilustración 17: Diagrama de Capas	65
Ilustración 18: Diagrama de Paquetes del Sistema	66
Ilustración 19: Diagrama Modelo Lógico (Entidad - Relación)	67
Ilustración 20: Diagrama Modelo Físico	68
Ilustración 21: Vista de casos de uso del sistema	69
Ilustración 22: Diagrama de Clases CUS, Transformar y Cargar Datos	76
Ilustración 23: Diagrama de secuencia CUS, Transformar y Cargar Datos	77
Ilustración 24: Menú Principal	78
Ilustración 25: Interfaz "Ejecución del proceso"	79
Ilustración 26: Ejecución correcta del ETL	80
Ilustración 27: Diagrama de Clases CUS, Consultar Información Cargada	83

Ilustración 28: Diagrama de secuencia CUS, Consultar Información Cargada	84
Ilustración 29: Interfaz “Consultar información cargada”	86
Ilustración 30: Información consolidado del alumno	87
Ilustración 31: Diagrama de clases CUS, Ejecutar Modelo	92
Ilustración 32: Diagrama de secuencia CUS, Ejecutar Modelo	93
Ilustración 33: Interfaz “Ejecución del Modelo Predictivo”	95
Ilustración 34: Consulta de resultados del modelo	96
Ilustración 35: Diagrama de Clases CUS, Consultar Reporte Predefinido	99
Ilustración 36: Diagrama de Secuencia CUS, Consultar Reporte Predefinido	100
Ilustración 37: Reporte Consolidado por año de ingreso	101
Ilustración 38: Reporte Detallado	102
Ilustración 39: Carga ETL, Integracion Services	104
Ilustración 40: Modelos de Predicción Evaluados	106
Ilustración 41: Leyenda del Gráfico de Elevación	107
Ilustración 42: Gráfico de elevación de minería de datos	108
Ilustración 43: Árbol de decisión de la deserción universitaria en la escuela informática	110
Ilustración 44: Error en la ejecución del Agente SQL	120
Ilustración 45: Agente SQL Server detenido	121
Ilustración 46: Agente SQL Server Iniciado	122
Ilustración 47: Error en el inicio de Servicio SQL	123
Ilustración 48: Servicio SQL Server Iniciado	123
Ilustración 49: Estrategia de Ejecución del modelo predictivo	127
Ilustración 50: Ingreso estimado por NSE en Lima Metropolitana	132
Ilustración 51: Distribución del Gasto por NSE en Lima Metropolitana	133
Ilustración 52: Variables del Modelo de Predicción	135
Ilustración 53: Correo electrónico de confirmación de ejecución del CUS: Transformar y Cargar datos	136

Índice de Tablas

Tabla 1: Estado de alumnos matriculados por semestre	5
Tabla 2: Definición de Deserción	17
Tabla 3: Tipología de la deserción	19
Tabla 4: Factores que impactan en la Deserción	20
Tabla 5: Descripción de entregables de la EDT	49
Tabla 6: Descripción del diagrama de proceso evaluación de deserción Universitaria	56
Tabla 7: Actores CUS, Transformar y Cargar datos	71
Tabla 8: Especificación CUS, Transformar y Cargar datos	71
Tabla 9: Actores CUS, Consultar información cargada	81
Tabla 10: Especificación CUS, Consultar información cargada	82
Tabla 11: Actores CUS, Ejecutar Modelo	88
Tabla 12: Especificación CUS, Ejecutar Modelo	89
Tabla 13: Actores CUS, Consultar reporte predefinido	97
Tabla 14: Especificación CUS, Consultar reporte predefinido	97
Tabla 15: Atributos usados en el modelo predictivo	124
Tabla 16: Cronograma del Proyecto	125
Tabla 17: Informe de Colección de Datos	128
Tabla 18: Informe de registros por variable	129
Tabla 19: Análisis de variables numéricas	130

INTRODUCCIÓN

La deserción es uno de los principales problemas que enfrentan las instituciones de Educación Superior. Es un problema preocupante, debido a las repercusiones sociales, institucionales y personales que trae como consecuencia. La deserción se puede definir como el proceso de abandono, voluntario o forzoso de la carrera en la que se matricula un estudiante, por la influencia positiva o negativa de circunstancias internas o externas a ella.

Según un estudio de la UNESCO del 2011, la tasa de deserción promedio de los países fue en primaria 8,3%, secundaria 15,5%, mientras que en las universidades la media latinoamericana que ronda el 50%. [Bibliografía Nro 17]

En América Latina el problema de la deserción universitaria no obedece a un solo factor; es un fenómeno que se explica a través de diversas variables, las cuales pueden o no estar relacionadas. Dichas variables pueden variar de acuerdo al contexto y las características personales, institucionales y económicas.

Otro estudio menciona que en 15 países que cubren un 90% de la Región, el costo de la deserción se estimaba en U\$ 11,1 billones de dólares al año, lo que en países como Brasil equivale al costo de dos millones de estudiantes universitarios. [Bibliografía Nro 16]

Los estudios revisados de investigaciones sobre la temática de la deserción identifican una variedad de causas, que van desde el abandono por la escasa formación previa, los reiterados fracasos en los exámenes finales, el origen social, la elección inadecuada de estudios, características familiares o circunstancias de la vida, problemas de organización de las diferentes unidades académicas, entre otras.

La deserción universitaria en el Perú alcanza una tasa del 17% del problema en educación, por consiguiente, la deserción implica una pérdida del capital de

recursos humanos para la familia, la comunidad y el país; conlleva, asimismo, a sentimientos de frustración que todo ser humano sufre al no lograr un objetivo.

El Ministerio de Educación del Perú propone el programa “Beca Doble Oportunidad” ofrece financiar los dos últimos años escolares pendientes y una carrera técnica; mientras que las universidades implementan estrategias que promuevan la gestión de conocimiento y la calidad de la formación.

La capacidad de monitorear e implementar intervenciones efectivas para disminuir la deserción por parte de las instituciones de educación superior es una de las estrategias que están adoptando algunos países diseñando un instrumento que permite la observación permanente de un estudiante cuando esté matriculado para identificar a aquéllos que presentan una mayor probabilidad de suspender temporal o definitivamente sus estudios, tomar medidas preventivas y mantener una información más precisa sobre la deserción estudiantil.

Antecedentes del Problema

En el Perú, un estudio del 2012 de la consultora Logros señala que entre 40 y 50 mil jóvenes abandonan sus estudios universitarios cada año, lo que representa no menos de cien millones de dólares desperdiciados por los padres de familia, con la consecuente frustración que ello representa para ellos mismos y sus hijos. De este grupo, el 70% corresponde a estudiantes de universidades privadas y el 30% a universidades estatales (Plasencia, 2011). Según Logros, se deben básicamente a dos factores: vocacional y económico.

Otro estudio del 2013 la consultora Flanqueo señala que las universidades orientadas al segmento socioeconómico C registran la mayor deserción con 28%, seguidas de las universidades del segmento B con 21% de deserción y las

universidades del segmento A con 15% de deserción. Los institutos muestran un panorama similar en el segmento socioeconómico C, donde la deserción estudiantil es del orden del 31%; en el B es de 20% y en el A, 13%. La deserción, además, puede implicar suspensión definitiva o temporal, voluntaria o forzada, lo cual depende de las razones: si es abandono de la carrera, abandono de la institución o abandono de la educación superior.

La principal razón para la deserción sería el factor económico, que representa el 29%; seguida por la falta de vocación (26%), decepción de la carrera elegida (20%), mala enseñanza (15%), mala organización y la falta de preocupación por el alumno (10%).

La deserción universitaria no obedece a un solo factor; es un fenómeno que se explica a través de diversas variables, las cuales pueden o no estar relacionadas. Diversos autores, como Escalante (2005), Merlino, et al. (2011), Salcedo, (2010), Sanabria (2002), Ramos (2003) y Rodríguez & Hernández (2008), respaldan que dichas variables pueden variar de acuerdo al contexto y las características personales, institucionales y económicas.

De la información revisada, en la ilustración 1 se ha elaborado el siguiente diagrama de Ishikawa, el cual nos muestra las diferentes características que influyen como causas en la deserción académica, estas características son agrupadas en tres grandes factores: Académico, Personal, Socio-Económico.

Las causas que podemos observar en el diagrama de Ishikawa nos ayudaran a definir la información que se necesitaran de los alumnos de la carrera de informática, se podrá analizar esta información y definir las variables que necesitaremos en nuestro modelo de predicción de deserción.

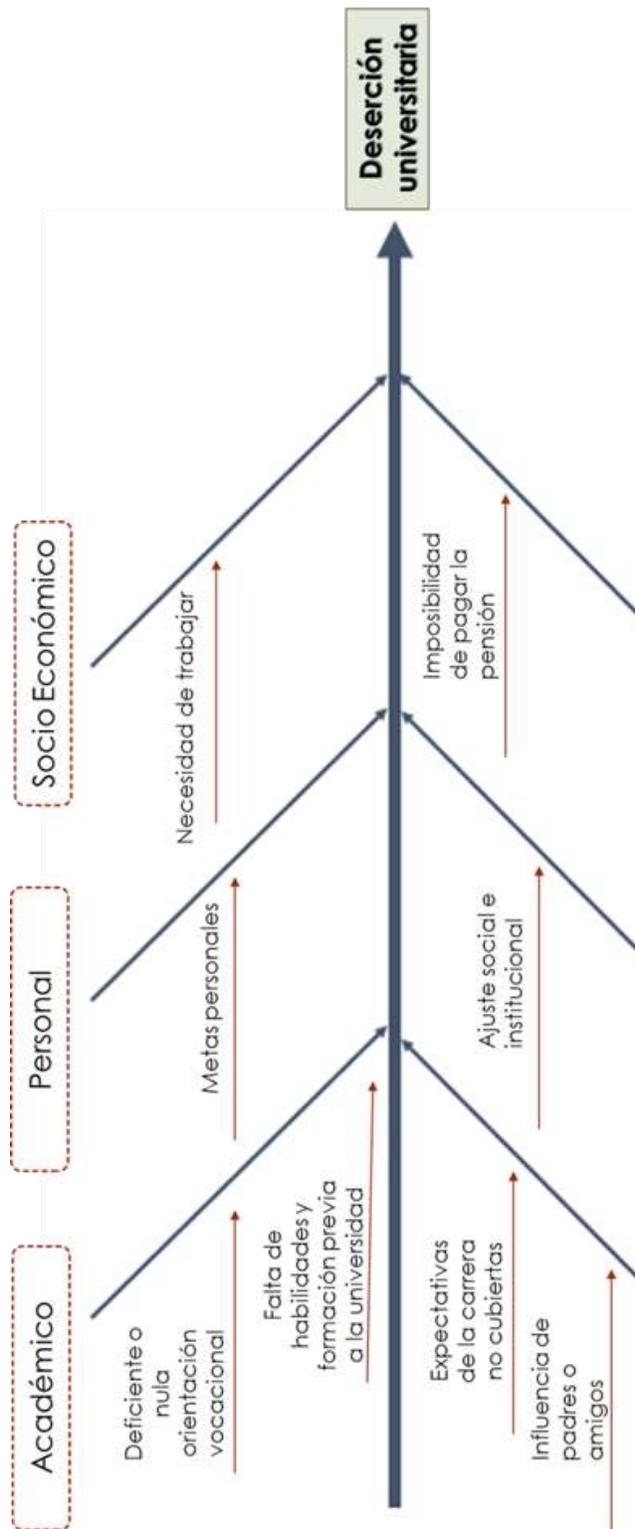


Ilustración 1: Diagrama de Ishikawa Causas de la deserción universitaria

En resumen, 46% de estudiantes inicia sus estudios sin estar suficientemente informado sobre la carrera, esto se debe a múltiples factores, tales como escoger la carrera de moda, falta de madurez, muchos estudiantes terminan la secundaria a los 15 o 16 años, falta de apoyo vocacional en el colegio porque pocas instituciones cuentan con un profesional orientado a este tema, y la elección de carreras universitarias no requeridas por el mercado. La facultad de ingeniería informática de la Universidad Ricardo Palma, no es ajena a esta realidad y en la siguiente tabla 1, presentamos el número de alumnos que se han matriculado por semestre y la cantidad de desertores por semestre.

La muestra tomada es del semestre 2010-I, hasta el semestre 2015 – I.

Tabla 1: Estado de alumnos matriculados por semestre

Semestre Académico	Nro. de Alumnos			
	Activos	Desertores	Terminaron	Matriculados
20101	28	25	12	65
20102	13	17	1	31
20111	38	40	1	79
20112	17	12	0	29
20121	31	31	0	62
20122	8	7	0	15
20131	33	18	0	51
20132	22	4	0	26
20141	40	7	0	47
20142	21	0	0	21
20151	42	0	0	42
Total general	293	161	14	468
%	63%	34%	3%	100%

Como podemos apreciar en la ilustración 2, la cantidad de desertores es alta lo cual nos da una vista del efecto de la deserción en la escuela informática, que en los últimos 4 años tiene una tasa promedio de 34%.

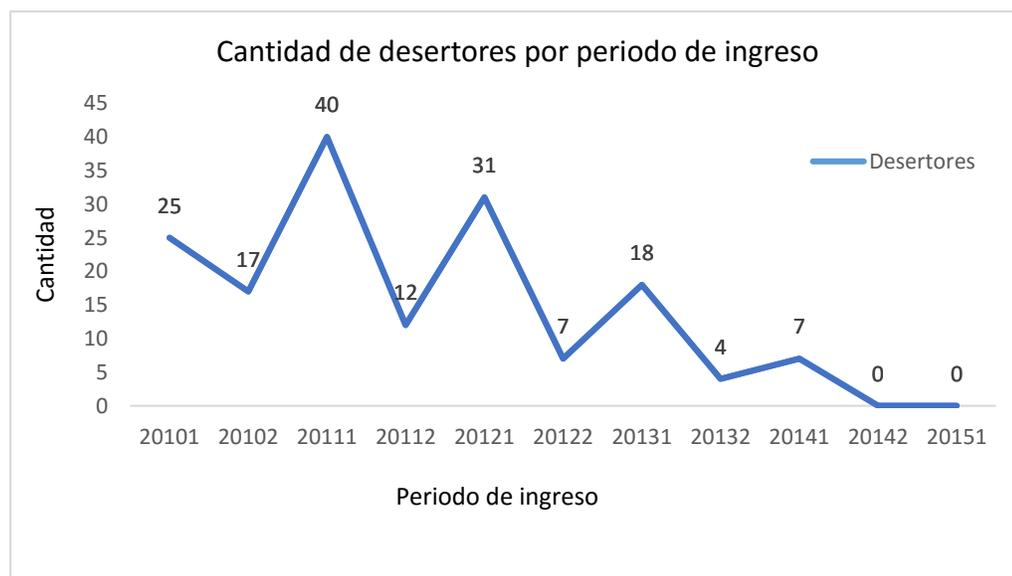


Ilustración 2: Alumnos desertores por periodo matriculado

La consecuencia del problema de la deserción es la pérdida de capital humano y una gran frustración por no lograr un objetivo, fundamentalmente para aquellos que salen del sistema de la educación superior.

Definición del Problema

La deserción universitaria es un fenómeno que se da en todo el Perú y la Universidad Ricardo Palma no es ajena a esta realidad.

Problema Central:

Deserción Universitaria de los alumnos de la escuela de Informática.

Problema específico:

1. No se cuenta con información histórica de deserción de los alumnos de la escuela de informática.
2. No se identifican a los alumnos con altas probabilidades de abandonar la carrera de Ingeniería Informática.
3. No se tienen reportes con indicadores de los alumnos con altas probabilidades de abandonar la carrera de Ingeniería Informática.

Propósito del Proyecto

A continuación explicaremos el objetivo general y los objetivos específicos del proyecto que se deben cumplir.

Objetivo General

Implementar un modelo predictivo de deserción estudiantil de pregrado en la escuela ingeniería informática de la Universidad Ricardo Palma.

Objetivos Específicos

1. Utilizar la información de los alumnos entregada por OFICIC y guardar la serie histórica de la deserción estudiantil, con la metodología definida en este estudio.
2. Identificar a los alumnos con altas probabilidades de abandonar la carrera de ingeniería informática con la finalidad de suministrar la

información obtenida a los responsables de la gestión académica y administrativa.

3. Explicar las causas encontradas de la deserción estudiantil en la escuela de informática y los indicadores que se utilizarán para su medición.
4. Mostrar los indicadores de la deserción a través de reportes predeterminados con el detalle de los alumnos que abandonan la carrera.

En la ilustración 3, podemos observar el diagrama de ubicación del proyecto.

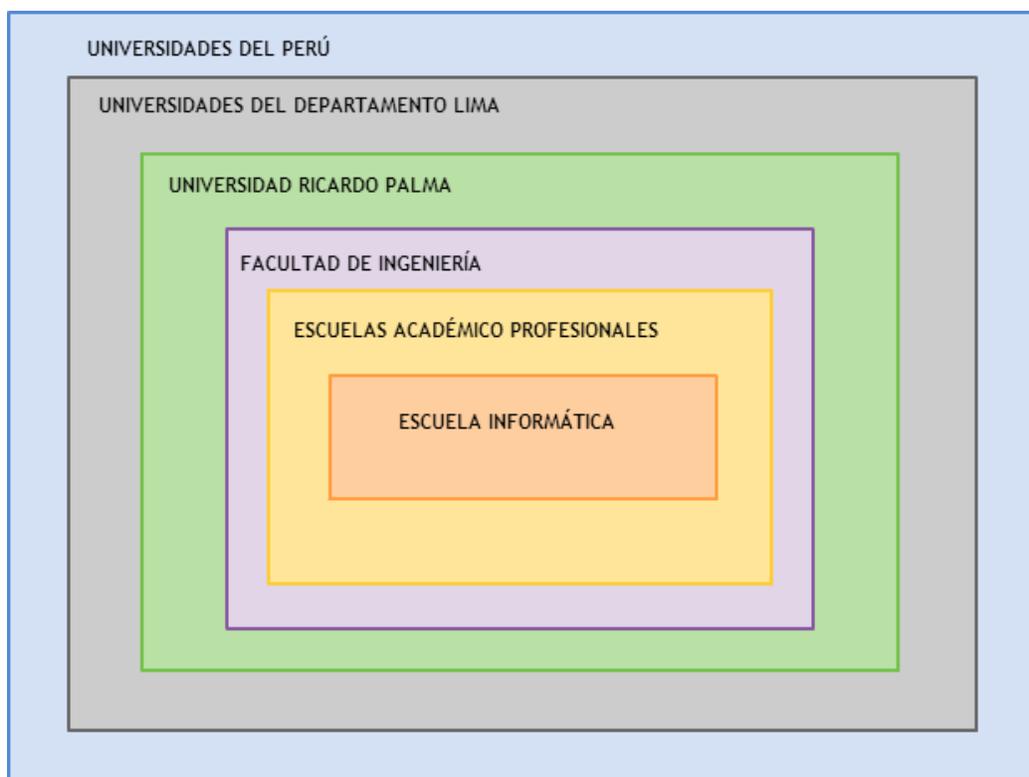


Ilustración 3: Diagrama de Ubicación del proyecto

Importancia del Proyecto

El proyecto permitirá a la Escuela de Informática a identificar a los alumnos con mayores probabilidades de desertar, de tal forma que pueda tomar acciones que prevengan el abandono de la carrera.

Con una buena gestión sobre la información proporcionada se podrían reducir los porcentajes de deserción en la escuela informática y por lo tanto aumentar el índice de egresados.

Justificación del Proyecto

Es preocupante observar que en la escuela de Ingeniería Informática de la Universidad Ricardo Palma, el 34% de estudiantes abandonan sus estudios superiores a raíz de diversos factores que actualmente para la universidad son desconocidos, causando graves efectos financieros, académicos y sociales, razón suficiente para analizar los resultados del presente estudio y plantear una solución.

Si se tomaran las estrategias necesarias para disminuir dichas deserciones, como consecuencia positiva la Universidad Ricardo Palma reduciría vacíos financieros y mejoraría la imagen y prestigio de la escuela de ingeniería informática.

La deserción de los alumnos cuestiona en muchos casos la calidad de enseñanza en los siguientes aspectos:

1. Formación de alumnos
2. Pasividad de los cursos
3. Dedicación de los docentes
4. Condiciones de estudio

Basándonos en programas contra la deserción universitaria realizados en Colombia por el Ministerio de Educación y en Honduras por la Universidad Autónoma de Honduras se tiene precedente de poder disminuir la deserción hasta en un 5%.

Definición de Términos

1. Deserción Universitaria

Tinto (1982) define deserción como una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo. Luego, es posible considerar como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante tres semestres académicos consecutivos. En algunas investigaciones, a este comportamiento se le denomina 'primera deserción' (first drop-out), ya que no se puede determinar si pasado este periodo de tiempo el individuo retomará o no sus estudios o si decidirá iniciar otro programa académico (Tinto 1989; Cabrera et al. 1992 y 1993; Adelman 1999).

2. Programa Académico

Un programa académico es el conjunto de cursos básicos, profesionales y complementarios, y actividades teóricas, prácticas y teórico prácticas integradas armónicamente mediante la interrelación de profesores, alumnos y recursos instrumentales tendientes a lograr una formación en determinadas áreas del conocimiento y a la obtención de un título académico.

3. Indicadores

Es una medida del nivel del desempeño de un proceso; el valor del indicador está directamente relacionado con un objetivo fijado de antemano. Normalmente se expresa en porcentaje.

4. Estadística Descriptiva

Es una gran parte de la estadística que se dedica a recolectar, ordenar, analizar y representar a un conjunto de datos, con el fin de describir apropiadamente las características de este.

5. Probabilidad

Es un método por el cual se obtiene la frecuencia de un acontecimiento determinado mediante la realización de un experimento aleatorio, del que se conocen todos los resultados posibles, bajo condiciones suficientemente estables.

6. Modelo Predictivo

Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos.

7. ETL

Proceso de extracción, transformación y carga de datos desde una fuente determinada hacia un data mart o el data warehouse corporativo.

8. Data Mart

Subconjunto del Data Warehouse que está orientado a un área específica del negocio.

Todas sus métricas y dimensiones están relacionadas con un área de negocio en particular.

9. Data Mining

Es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.

El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

10. EDT

El propósito de una EDT es organizar y definir el alcance total aprobado del proyecto.

Su forma jerárquica permite una fácil identificación de los elementos finales llamados "Paquetes de Trabajo", será ejecutado por el equipo de proyecto para cumplir con los objetivos de éste y crear los entregables requeridos.

La EDT, sirve como la base para la planificación del proyecto.

Resumen

El problema de deserción estudiantil afecta a todas las entidades educativas de Latinoamérica, siendo los índices más alarmantes en la deserción universitaria.

Las causas de la deserción pueden darse por factores personales, académicos o socio económicos.

En la Universidad Ricardo Palma, en la facultad de Ingeniería escuela Informática existe deserción estudiantil que repercute principalmente en el prestigio de la entidad. Un modelo predictivo ayudaría a identificar los perfiles de alumnos desertores y así tomar acciones frente a los alumnos propensos a abandonar la universidad.

CAPITULO I: MARCO TEORICO

Introducción

En este capítulo se presenta la metodología y técnicas que permitió desarrollar el presente trabajo. Se muestran aspectos como el tipo de investigación, las técnicas y procedimientos que fueron utilizados para llevar a cabo dicha investigación.

Se revisaron diferentes fuentes de información como papers, tesis y estudios que nos permitieron entender mejor la deserción universitaria.

Revisión de la literatura

1. Tesis revisadas

A continuación se citan las diversas tesis revisadas y el estudio que se realizaron en ellas.

1.1 La deserción estudiantil en la facultad de ciencias económicas y empresariales de la universidad de Piura y su impacto en los resultados económicos. [Bibliografía Nro 6]

Se realizó un extenso estudio sobre el grado de deserción universitaria en la Facultad de Ciencias Económicas y Empresariales de la Universidad de Piura.

El análisis realizado es un trabajo exploratorio, de tipo longitudinal, en el cual se utilizó estadística descriptiva y el Programa IBM SPSS Statistics versión 21.0 para tratar los datos de los alumnos que desertaron los estudios en la Facultad de Ciencias Económicas y Empresariales.

Se realizó la Prueba de Independencia Chi Cuadrado la cual permitió determinar si existe una relación entre dos variables categóricas.

Se considera que esta problemática se podría disminuir haciendo mejores esfuerzos en el tema del Asesoramiento Académico.

1.2 Modelos de análisis de la deserción estudiantil en la educación superior (Chile). [Bibliografía Nro 3]

Este artículo aborda el tema de la deserción y retención de estudiantes en la educación superior desde una perspectiva conceptual.

Se revisan diferentes enfoques teóricos que han sido empleados en esta investigación, estos modelos fueron desarrollados enfatizando factores psicológicos, económicos, sociológicos, organizacionales o aspectos entre la iteraciones entre el estudiante y la institución. Dichos factores resultan predictivos del abandono.

Se identificaron tres criterios para la selección. Estos fueron los siguientes:

1. Estudiantes que dejaron de estudiar habiéndose matriculado.
2. Estudiantes que no se matricularon en el año académico.
3. Estudiantes que retiraron sus documentos luego de matricularse.

1.3 Deserción universitaria en estudiantes de una universidad privada de Iquitos. [Bibliografía Nro 5]

Se realizó una investigación cualitativa de tipo exploratorio, debido a la necesidad de comprender en profundidad las razones de la deserción universitaria en jóvenes que abandonaron los estudios en una universidad privada de la ciudad de Iquitos.

Se analizan los resultados obtenidos del estudio y se plantea alternativas de solución, entre ellas: “Los profesores se integren a los objetivos que persigue la institución”, “Promover espacios de diálogo entre la familia universitaria”, “Mejorar el proceso enseñanza-aprendizaje por parte de los docentes con metodologías, talleres y herramientas aportados por la universidad de Piura.

1.4 Deserción y retención estudiantil en los programas de pregrado de la pontificia universidad javeriana (Bogotá D.C., noviembre de 2009). [Bibliografía Nro 4]

El estudio presenta un análisis de las variables y causas del fenómeno de la deserción en la Pontificia Universidad Javeriana, a partir del cual se propone un plan de retención.

Se define como desertor “aquel estudiante que abandona la institución educativa durante dos periodos consecutivos, como resultado de la interacción o del efecto individual y combinado de diferentes categorías de variables individuales, académicas, institucionales y socioeconómicas”

Entre las categorías de variables se destacan el género, la edad, el estado laboral, el ingreso económico, la vivienda propia, el número de hermanos, el nivel educativo de la madre, así como también algunas relacionadas con el entorno socioeconómico, como la tasa de desempleo, el tipo de la institución y de programa, los apoyos financieros y académicos.

2. Deserción Universitaria

Según el Diccionario de la Real Academia Española, “desertar” significa “abandonar las obligaciones o los ideales”, “separarse o abandonar la causa o apelación”. En nuestro estudio concretamente nos referimos a la deserción universitaria.

En la Tabla 2, observamos algunas definiciones de deserción en el ámbito educativo.

Tabla 2: Definición de Deserción

Año	Autor	Definición de Deserción Estudiantil
2000	Sposetti	Proceso de selección que se opera en la Enseñanza Superior, una medida del rendimiento académico del alumnado y de la Eficacia del Sistema Educativo en general.
2003	Ramos	Proceso de abandono, voluntario o forzado, de la carrera en la que se matricula un estudiante, por la influencia positiva o negativa de circunstancias internas o externas a él o ella.
2005	Himmel	Abandono prematuro de un programa de estudios antes de alcanzar el grado o título.

2008	Rodríguez y Hernández	Disolución del vínculo estipulado a través de la matrícula académica, ya sea por parte del estudiante o la universidad. Ésta tiene graves efectos financieros, académicos y sociales para ambos.
2008	Rojas y Gonzales	Tipo de conducta con la cual el estudiante decide o se ve forzado a abandonar sus estudios.

Basándonos en los conceptos anteriores, podemos definir qué DESERCIÓN es la suspensión definitiva y voluntaria, que se puede distinguir por diferentes modalidades, tales como: abandono de la carrera, abandono de la institución y el abandono del sistema de educación superior.

2.1 La deserción desde el punto de vista institucional

La pérdida de estudiantes causa serios problemas financieros a las instituciones al producir inestabilidad en la fuente de sus ingresos.

Esto es en particular evidente en el sector privado, en el que las colegiaturas constituyen parte sustancial de los ingresos institucionales, pero no es menos importante en el sector público debido a los presupuestos insuficientes.

La dificultad que afrontan las universidades para definir la deserción, consiste en identificar qué tipos de abandono, la decisión de abandonar puede obedecer a distintas causas; algunas de ellas son susceptibles de intervención institucional, otras no. [Bibliografía Nro 7]

2.2 Tipología

Luego de haber definido el concepto de Deserción con el cual vamos a trabajar durante el desarrollo de esta investigación, en este apartado vamos a explicar los diversos tipos de Deserción de acuerdo a los estudios de Vincent Tinto (1993), los cuales explica más en detalle en la Tabla 3.

Tabla 3: Tipología de la deserción

Deserción	Total	Facultad	Escuela	A Primer semestre de la carrera
Definición	Abandono definitivo de la formación académica Individual.	Cambio de Facultad	Cambio de Escuela en una misma Facultad.	Por la inadecuada adaptación a la carrera Universitaria.

2.3 Variables Asociadas a la Deserción

En múltiples ocasiones, las causas por las cuales un estudiante deserta son muy difíciles de pronosticar; sin embargo, la deserción se ha asociado usualmente a variables como problemas económicos, mala elección de la carrera, falta de interés por el desarrollo personal y profesional, entre otras.

En la tabla 4, se clasifican los factores de deserción por el nivel de impacto, Alta, media o baja.

Tabla 4: Factores que impactan en la Deserción

Intensidad	Factores
Alta	<p>Factores económicos que impiden la continuidad del desertor en la Universidad.</p> <p>Bajos niveles de comprensión, unidos a la falta de interés y apatía por programas curriculares.</p> <p>Cursos no asociados ni aplicables con el ejercicio profesional.</p> <p>Orientación profesional.</p>
Media	<p>Ambientes educativos universitarios en los cuales está inmerso el estudiante.</p> <p>Proceso educativo y acompañamiento al estudiante en su formación.</p> <p>Adaptación social del estudiante desertor con sus pares u homólogos.</p> <p>Programas micro curriculares universitarios rígidos con respecto a los de su formación secundaria, de alta intensidad temática, dispuestos en corto tiempo.</p> <p>Modelos pedagógicos universitarios diferentes a los modelos de bachillerato, que imprime un alto nivel de exigencia.</p> <p>Evaluaciones extenuantes y avasalladoras. Las evaluaciones y trabajos universitarios tienen mucho nivel de complejidad que las de secundaria.</p>
Baja	<p>Ambientes familiares.</p> <p>Edad. La mayoría de los estudiantes universitarios son muy jóvenes.</p> <p>Cantidad de oferentes.</p> <p>Masificación de la educación.</p>

2.4 Perfil del desertor

Abordar el estudio de la deserción estudiantil universitaria, conduce a la identificación de las características de los estudiantes que desertan de su formación, en cualquier nivel de educación superior. Esto constituye el perfil del desertor, que es lo que vamos a analizar en este apartado.

Según Himmel (2005), las personas que desertan, en cualquier institución educativa, bajo cualquier circunstancia, presentan en mayor o menor grado, algunas de las siguientes características:

1. Problemas de disciplina

Normalmente los estudiantes que dejan sus estudios son los que faltan o han faltado a las normas del centro educativo en particular. Estos alumnos son los que generan más costos, en función de tiempo, para la organización.

2. Nivel socioeconómico bajo o sin opción económica

Sin duda alguna, es muchas veces una de las características que más predomina en las personas que desertan de alguna institución educativa.

3. Ausentismo de clases

Algunas veces por trabajo u otras labores; algunas otras por desinterés en su formación académica.

4. Problemas de salud psicosomática

Problemas psicológicos que se demuestran a través de lo corpóreo.

5. Problemas inherentes a la edad

La rebeldía de la adolescencia puede influir en la decisión de abandonar los estudios.

6. Inadecuadas relaciones interpersonales

Estudiantes que presentan dificultades para adaptarse al ambiente universitario, se les hace muy difícil construir lazos de amistad y muchas veces son introvertidos.

7. Resistencia a desarrollar actividades formativas

Participación de congresos, confraternidades de la Facultad en la que estudian, dinámicas en equipo, entre otras.

8. Inapetencia por el conocimiento

Estudiantes que desertan en los estudios porque no está inherente en ellos el deseo de estudiar.

9. Desmotivación hacia la carrera y/o a la universidad

La mala elección de la carrera universitaria es uno de los factores que más realzan en los desertores.

2.5 Actores que intervienen en la deserción estudiantil universitaria

En este apartado se identifican a las personas que intervienen, tanto directa como indirectamente, en el proceso de deserción de cualquier estudiante universitario. Se involucran, en el fenómeno de la deserción, los siguientes actores:

1. Desertores

Son los estudiantes que se han retirado de su formación, en este caso universitaria.

2. Padres de Familia de los alumnos desertores.

Los padres de familia, muchas veces por motivos económicos, hacen que los alumnos se vean obligados a dejar los estudios en cierta universidad, para poder continuarlos en algún instituto o en otra universidad; en otros casos, la precaria economía familiar podría forzar a los alumnos a dejar por completo los estudios universitarios y buscar un trabajo para contribuir en la mejora de la misma.

3. Ex compañeros de estudio del semestre del cual se retiró el desertor.

La manera en la que los ex compañeros pueden influir es mediante las opiniones que tienen sobre la Facultad en la que estudian, los profesores, las asignaturas que se dictan; así como las demás universidades en las que pueden estudiar.

4. Profesores, Directivos y administradores académicos, quienes acompañaron al desertor en el proceso educativo hasta el semestre de retiro.

En el caso de los profesores, directivos y administradores académicos, muchas veces su intervención suele presentarse en la forma en la que se relacionan con los alumnos, más que por el grado de dificultad de la asignatura.

La mayoría de los alumnos que se retiran, lo hacen porque no logran relacionarse adecuadamente con su asesor académico, o porque percibe que no recibe la suficiente ayuda por parte de las autoridades académicas.

Cada uno de los actores mencionados anteriormente, intervienen de manera directa o indirecta, en la decisión de un alumno de desertar con sus estudios universitarios.

3. Metodologías

3.1 Rational Unified Process o Proceso Unificado de Rational (RUP)

RUP es una metodología de desarrollo de software y junto con el Lenguaje Unificado de Modelado (UML), constituye la metodología estándar más utilizada para el análisis, implementación y documentación de sistemas orientados a objetos.

Lenguaje Unificado de Modelado (UML), es el lenguaje de modelado de sistemas de software más conocido y utilizado en la actualidad; está respaldado por el OMG. Es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema. UML ofrece un estándar

para describir un "plano" del sistema (modelo), incluyendo aspectos conceptuales tales como procesos de negocio y funciones del sistema, y aspectos concretos como expresiones de lenguajes de programación, esquemas de bases de datos y componentes reutilizables.

Los autores de RUP destacan que el proceso de software propuesto por RUP tiene tres características esenciales:

1. Está dirigido por los Casos de Uso.
2. Está centrado en la arquitectura.
3. Es iterativo e incremental.

3.2 Metodologías para la minería de datos.

En muchos proyectos implementan el proceso KDD, mientras que otras aplican un estándar más específico como CRISP-DM. Para trabajar con productos de la empresa SAS, tiene a su disposición una metodología especialmente desarrollada para los mismos, la metodología SEMMA.

Por otro lado, la metodología Catalyst (conocida como P3TQ) está ganando cada vez mayor popularidad debido a su completitud y flexibilidad para adaptarse en distintos escenarios.

En la ilustración 4 podemos observar la encuesta realizada por KDnuggets en el año 2007, el cual nos muestra que la metodología más usada es la minería de datos es CRIPS-DM en un 42%.

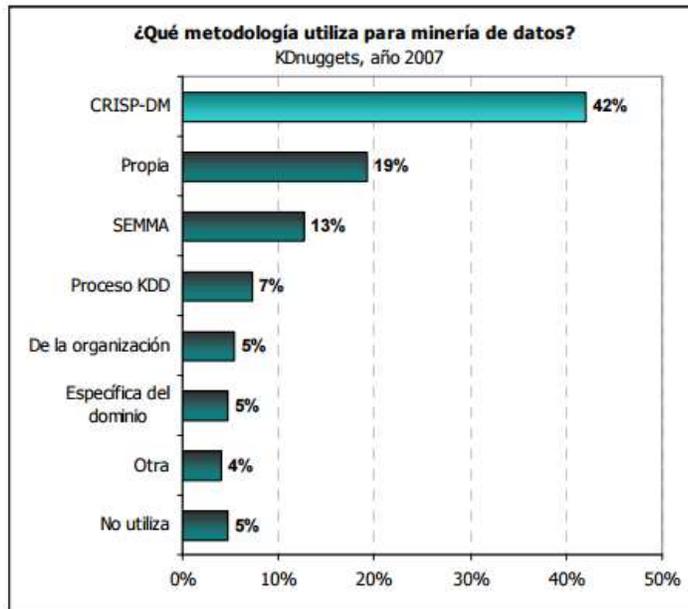


Ilustración 4: Encuesta realizada por la KDnuggets en el año 2007

1. Metodología CRISP-DM

La metodología CRISP-DM consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. [Bibliografía Nro 9]

En la ilustración 5, podemos observar los cuatro niveles de la metodología. A nivel más general, el proceso está organizado en fases, estando cada fase a su vez estructurada en varias tareas genéricas de segundo nivel.

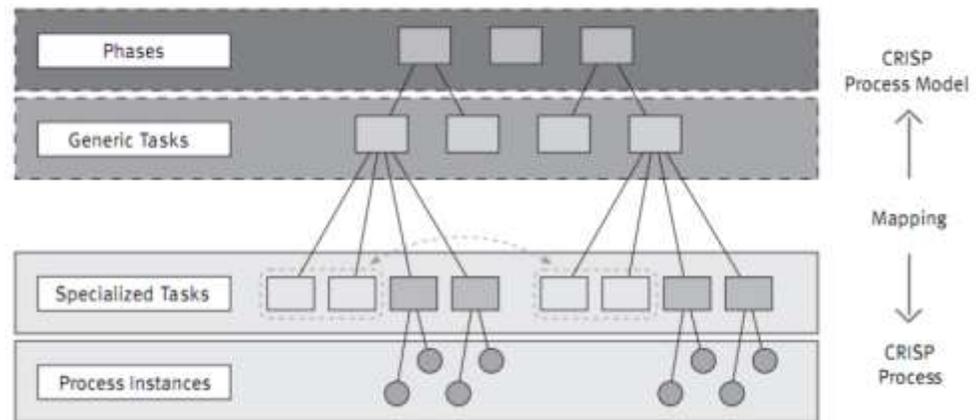


Ilustración 5: Cuatro niveles de detalle de la metodología CRISP-DM

Las tareas genéricas se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea genérica “limpieza de datos”, en el tercer nivel se indican las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, “limpieza de datos numéricos”, o “limpieza de datos categóricos”. El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de Data Mining específico.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

1. Comprensión del negocio

En esta etapa se establecen los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito).

Se evalúa la situación, es decir los recursos, requerimientos y supuestos del negocio. Se establecen los objetivos de la minería de datos y el plan del proyecto.

2. Comprensión de los datos

Un proyecto de minería de datos con éxito parte de una pregunta o necesidad bien definida.

Esta fase se inicia con la obtención de los datos. Una vez conseguida la información se procede a familiarizarse con ella e identificar su procedencia.

3. Preparación de datos

Hernández Orallo en [Hernández Orallo et al., 2004] escribió:

“La calidad del conocimiento descubierto no solo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados. Por ello, después de la recopilación, el siguiente paso en el proceso es seleccionar y preparar el subconjunto de datos que se van a minar, los cuales constituyen lo que se conoce como vista minable”

Estos datos son de una o más fuentes de datos que se utilizará para la exploración y el modelado.

Preparación de los datos es a menudo un proceso que consume tiempo y fuertemente propenso a errores. El viejo dicho "basura in-basura out" es particularmente aplicable a los proyectos de minería de datos donde los datos que se

reunieron, muchos de ellos están fuera de rango, no válidos o faltan valores.

Entonces, el éxito de los proyectos de minería de datos depende en gran medida de la calidad de los datos preparados.

Las variables sirven como marcadores de posición para los datos. Hay dos tipos de variables, numéricos y categóricos.

1. Numérica o variable continua, puede aceptar cualquier valor dentro de un intervalo finito o infinito, por ejemplo, altura, peso, etc.
2. Categórica o variable discreta, puede aceptar dos o más valores (categorías). Hay dos tipos de datos categóricos: Nominal y Ordinal.

Los datos nominales no tienen un orden intrínseco en las categorías. Por ejemplo, "género" con dos categorías, masculino y femenino.

Por el contrario, los datos ordinales tienen una ordenación intrínseca en las categorías. Por ejemplo, "nivel de energía" con tres categorías ordenadas (baja, media y alta).

En la ilustración 6, podemos observar como están clasificados los datos y los tipos de variables que usan en el proceso de minería de datos.

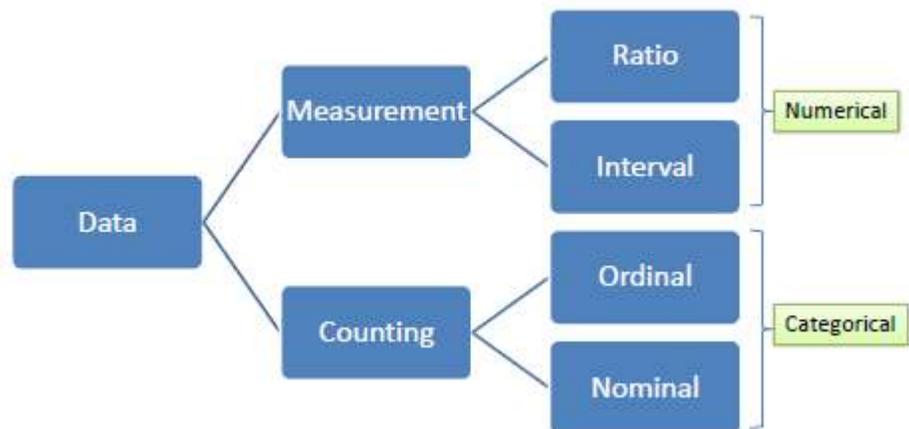


Ilustración 6: Categoría de las variables

4. Modelado

Aquí es donde se produce conocimiento nuevo, construyendo modelos basados en los datos recopilados. El modelo describe los patrones y relaciones existentes en los datos. Estos patrones y relaciones son los que se pueden utilizar para entender mejor los datos, predecir comportamientos o explicar situaciones observadas. En esta etapa se deben tomar las siguientes decisiones:

1. Elegir la tarea de Data Mining apropiada para el objetivo del proyecto y para los datos involucrados. Por ejemplo, se puede decidir usar una tarea descriptiva que ayude a conocer con más precisión las características de los alumnos desertores de la Universidad.
2. Elegir el tipo de modelo, por ejemplo se puede elegir el agrupamiento para obtener grupos de alumnos con

características semejantes que puedan ser descriptos apropiadamente.

3. Elegir el algoritmo de Data Mining que resuelva la tarea y ofrezca un modelo resultante. Para tomar esta determinación, en capítulos subsiguientes se describen en detalle los algoritmos plausibles de ser seleccionados para el objetivo planteado.
4. Las tareas pueden ser predictivas o descriptivas. Dentro de las tareas predictivas se encuentran la clasificación y la regresión. Son tareas descriptivas el agrupamiento, las reglas de asociación y las correlaciones.

5. Evaluación

Modelo de Evaluación es una parte integral del proceso de desarrollo del modelo. Ayuda a encontrar el mejor modelo que representa nuestros datos y lo bien que el modelo elegido trabajará en el futuro.

Evaluación del modelo se puede dividir a dos secciones:

1. Evaluación de la clasificación
2. Evaluación de regresión

6. Despliegue del modelo

Normalmente los proyectos de Data Mining no terminan en la implantación del modelo, sino que se deben documentar y presentar los resultados de manera comprensible. Dependiendo de los requisitos, la fase de despliegue puede ser tan simple como la generación de un informe o tan complejo como la implementación de un proceso de minería de datos repetibles.

En la ilustración 7, podemos observar todo el ciclo completo del proceso de la metodología CRISP DM, empezando primero por el análisis del problema hasta terminar con la última fase de explotación del resultado

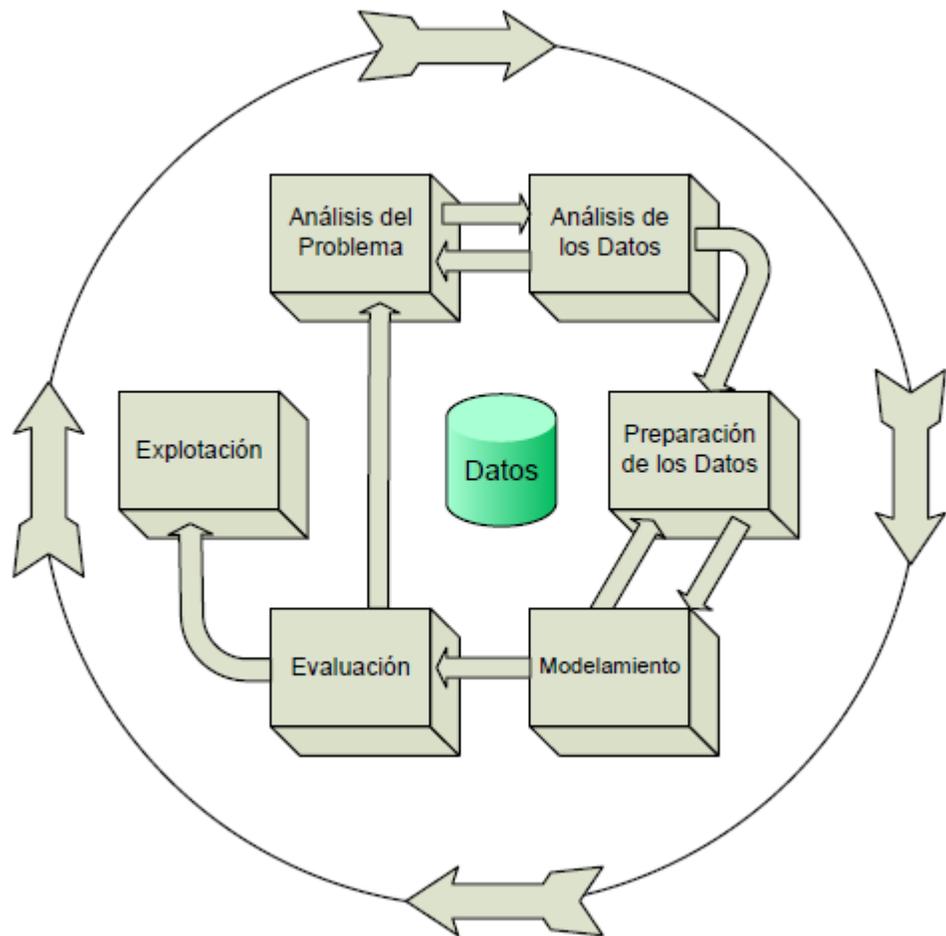


Ilustración 7: El proceso CRISP DM

4. Data Mining (Minería de datos)

Se define la Minería de Datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

Otra definición: es el análisis de archivos y bitácoras de transacciones, trabaja a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones.

“La minería de datos es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Es capaz de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido de la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades. Permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en que deben actuar y tomar decisiones.” [Hernández Orallo et al., 2004]

El campo de la Minería de Datos es relativamente nuevo y está en estado de evolución. La primera conferencia internacional de KDD (Knowledge Discovery in Databases) y DM (Data Mining), se llevó a cabo en 1995 y hay una variedad de definiciones para este concepto.

Es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos.

4.1 Técnicas de minería de datos

Las técnicas de minería de datos apuntan a transformar datos en información para la toma de decisiones. Dependen en gran medida de los datos de que se disponga y de la preparación adecuada que se les dé a los mismos, de manera de poder utilizar diferentes algoritmos y metodologías de descubrimiento.

El objetivo de las técnicas de minería de datos es extraer conocimiento desde los datos, y ese conocimiento constituye el modelo de los datos analizados. Los patrones pueden ser utilizados para predecir observaciones futuras o explicar observaciones pasadas, capacidades

fundamentales para mejorar el comportamiento en relación a un fenómeno, como el caso de la deserción universitaria. Si bien las técnicas de minería de datos ocupan un lugar relevante en la empresa y la toma de decisiones del sector privado, también son aplicables a la educación superior, considerando las grandes cantidades de datos que conforman el expediente de los estudiantes.

Con dicha información las Universidades podrían conocer los perfiles de sus alumnos, así como las características de los estudiantes desertores. [Bibliografía Nro 2]

Los algoritmos de minería de datos se clasifican en dos grandes categorías:

1. Exploración del pasado, explica el pasado a través de la exploración de datos.
2. Predicción del futuro, predice el futuro a través del Modelado.
Arboles de decisión

a) Árbol de decisión

Los arboles de decisión son una técnica de minería de datos, que establece un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. [Bibliografía Nro 8]

Algoritmo ID3

El algoritmo ID3 es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos.

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, (el atributo a clasificar) es el objetivo, el cual es de tipo binario (positivo o negativo, sí o no, válido o inválido, etc.).

El algoritmo ID3 se apoya en técnicas matemáticas y probabilísticas, es usado para ayudar a decidir qué atributo debe ser el siguiente en seleccionarse.

ID3 realiza esta labor mediante la construcción de un árbol de decisión. Los elementos son:

Nodos: Los cuales contendrán atributos.

Arcos: Los cuales contienen valores posibles del nodo padre.

Hojas: Nodos que clasifican el ejemplo como positivo o negativo.

Si todos los ejemplos son positivos devolver un nodo positivo.

Si todos los ejemplos son negativos devolver un nodo negativo.

Si Atributos está vacío devolver el voto mayoritario del valor del atributo objetivo.

Algoritmo C4.5

Algoritmo usado para generar un árbol de decisión desarrollado por Ross Quinlan. [Bibliografía Nro 1] C4.5 es una extensión del algoritmo ID3 desarrollado anteriormente por Quinlan. Los árboles de decisión generados por C4.5 pueden ser usados para clasificación.

C4.5 construye árboles de decisión desde un grupo de datos de entrenamiento de la misma forma en que lo hace ID3, usando el concepto de entropía de información.

En C4.5 se hicieron un número de mejoras a ID3. Algunas de ellas son:

Manejo de ambos atributos continuos y discretos, a fin de manejar atributos continuos, C4.5 crea un umbral y luego se divide la lista en aquellos cuyo valor de atributo es superior al umbral y los que son menores o iguales a él.[Bibliografía Nro 3]

Manejo de los datos de formación con valores de atributos faltantes C4.5, permite valores de los atributos para ser marcado como faltantes. Los valores faltantes de los atributos simplemente no se usan en los cálculos de la ganancia y la entropía.

Manejo de atributos con costos diferentes podando árboles después de la creación - C4.5, se remonta a través del árbol una vez que ha sido creado e intenta eliminar las ramas que no ayudan, reemplazándolos con los nodos de hoja.

b) Clustering

Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores que utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

La técnica consiste en agrupar un conjunto de datos, sin tener clases predefinidas, basándose en la similitud de los valores de los atributos de los distintos datos. Esta agrupación, a diferencia de la clasificación, se realiza de forma no supervisada, ya que no se conoce de antemano las clases del conjunto de datos de entrenamiento.

K-Means es un método particional de clustering donde se construye una partición de una base de datos D de n objetos en un conjunto de k grupos, buscando optimizar el criterio de particionamiento elegido. [Bibliografía Nro 6]

¿Por qué usar minería de datos?

1. Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios. contribuye a la toma de decisiones tácticas y estratégicas.
2. Proporciona poder de decisión a los usuarios del negocio, y es capaz de medir las acciones y resultados de la mejor forma.

3. Genera Modelos descriptivos: permite a empresas, explorar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.
4. Genera Modelos predictivos: permite que relaciones no descubiertas través del proceso de minería de datos sean expresadas como reglas de negocio.

4.2 Análisis predictivo

El Análisis predictivo utiliza estadística junto con algoritmos de minería de datos. Se basan en el análisis de los datos actuales e históricos para hacer predicciones sobre futuros eventos. Dichas predicciones raramente suelen ser afirmaciones absolutas, pareciéndose más a eventos y su probabilidad de que suceda en el futuro.

En el mundo de los negocios los modelos predictivos explotan los patrones de comportamiento encontrados en el pasado (Datos históricos) para poder identificar riesgos y oportunidades. Los modelos capturan las relaciones entre muchos factores permitiendo capturar riesgos potenciales asociados a un conjunto de condiciones, guiando así a la toma de decisiones. [Bibliografía Nro 1]

En la ilustración 8, podemos observar todas las técnicas de minería de datos que existen agrupadas en dos grandes grupos: Exploración del pasado y predicción del futuro, en nuestra tesis abarcaremos la rama de predicción del futuro para saber los posibles desertores de la carrera.



Ilustración 8: Mapa de la Minería de Datos [Bibliografía Nro 10]

5. Tecnologías

A continuación se citan las tecnologías que se revisaron y se usaron en nuestro proyecto.

1. Plataforma de desarrollo Visual Studio.

Visual Studio es una plataforma de desarrollo integrada de Microsoft que permite a los desarrolladores crear aplicaciones, sitios y aplicaciones web, así como servicios web en cualquier entorno que soporte la plataforma .NET (a partir de la versión net 2002). Así se pueden crear aplicaciones que se intercomunican entre estaciones de trabajo, páginas web y dispositivos móviles.

Visual Studio 2012 Update 2 (Visual Studio 2012.2) es la última actualización para Visual Studio, acompañada por .NET Framework 4.5, proporcionando nuevas funcionalidades y correcciones

2. Base de datos

Base de datos recoge, almacena y gestiona la información que los usuarios puedan recuperar, añadir, actualizar o eliminar dicha información. Se presenta información en tablas con filas y columnas.

Una tabla se conoce como una relación en el sentido de que es una colección de objetos del mismo tipo (filas).

Los datos en una tabla se pueden relacionar de acuerdo a las teclas comunes o conceptos, y la capacidad de recuperar datos relacionados de tablas relacionadas es la base de la base de datos relacional.

La mayoría de las cajas de herramientas de minería de datos se conectan a las bases de datos a través de ODBC (Open Database Connectivity) o JDBC (Java Database Connectivity) interfaces.

En la ilustración 9, se observa una representación de una base de datos.



Ilustración 9: Representación de una base de datos

Structured Query Language (SQL), es un lenguaje de programación de base de datos para la gestión y manipulación de datos en los sistemas de gestión de bases de datos relacionales (RDBMS).

Data Definition Language (DDL), permite crear, modificar, eliminar las tablas de las bases de datos. También podemos definir índices (claves), especificar vínculos entre tablas, e imponer restricciones entre las tablas de bases de datos.

CREATE TABLE: crea una nueva tabla

ALTER TABLE: altera una tabla

DROP TABLE: elimina una tabla

CREATE INDEX: crea un índice

DROP INDEX: elimina un índice

Data Manipulation Language (DML), es un lenguaje que permite a los usuarios acceder y manipular los datos.

SELECT: recuperación de datos de la base de datos

INSERT INTO: inserción de nuevos datos en la base de datos

ACTUALIZACIÓN: modificación de datos en la base de datos

BORRAR: supresión de los datos en la base de datos

3. Proceso Extraccion, Transformacion y Carga (ETL)

Extracción de los datos.

Proporciona la capacidad de extraer datos de una variedad de fuentes de datos, como archivos planos, bases de datos relacionales, datos de streaming, archivos XML, y fuentes de datos ODBC / JDBC.

La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

Transformación de datos

Proporciona la capacidad de limpiar, convertir, combinar los datos. La fase de transformación aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados

Carga de datos

La fase de carga es el momento en el cual los datos de la fase anterior (transformación) son cargados en el sistema de destino.

En la ilustración 10, podemos observar cómo se realiza el proceso ETL, empezando por la extracción de la información de una fuente origen, luego se transforma los datos y por último se carga a la fuente destino.

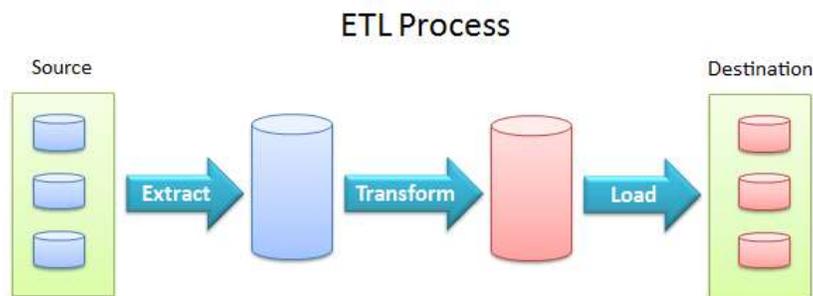


Ilustración 10: Proceso ETL

Conclusiones

Esta investigación es un estudio de tipo descriptivo - explicativo, que busca identificar las particularidades del fenómeno de la deserción en la Universidad Ricardo Palma y explicar por qué los estudiantes interrumpen sus estudios. Para esto, se emplea una metodología cuantitativa representada en el procesamiento de datos estadísticos. Asimismo, se complementa con la investigación cualitativa, debido a que se trató de determinar el porqué del comportamiento de los estudiantes que desertan los estudios universitarios, y qué variables son las que influyen en esta decisión.

Luego de revisar la literatura de la tesis se eligieron los modelos de Arboles de decisión y Clustering como los más adecuados para los objetivos de predicción que contempla la presente investigación.

Se realizó una revisión sobre el tema a nivel internación, nacional y por último se realizará a nivel institucional con el fin de conocer los índices, factores, causas y métodos de análisis de la deserción. Producto de dicha revisión se elaboró un estado del arte sobre el tema y el marco conceptual que delimita el alcance de esta investigación.

La recolección de los datos de los alumnos retirados durante los años 2010 hasta 2015 se obtendrá de la OFICIC de la Universidad Ricardo Palma.

CAPITULO II: DESARROLLO DEL PROYECTO

1. Alcance del Proyecto

El proyecto utilizó una muestra de la población total de alumnos de los últimos 4 años de la Escuela informática y luego de una limpieza de datos, fue objeto de investigación bajo técnicas de minería de datos que nos permitieron obtener un modelo que procesa la información y predice sus probabilidades de deserción.

1. Integration Services como herramienta BI de carga de información.
2. Analysis Services como herramienta BI para ejecución del modelo predictivo.
3. SQL Server como base de datos del proyecto y C# (.NET) como lenguaje de programación.
4. Se elegirá el modelo cuyo resultado sea más cercano al modelo ideal que propone la herramienta Analysis Services.

Alcance del producto

1. Base de datos final depurada sin repitencias o registros inválidos.
2. Se analizó la historia de 4 años en el modelo predictivo.
3. Modelo Analítico de minería de datos evaluado y bien estructurado que permita predecir la probabilidad de que un estudiante abandone sus estudios, dadas sus características socio económico y académico.

4. Interfaz de usuario para la ejecución del proceso de carga de información depurada a la base de datos del proyecto.
5. Interfaz de usuario para la consulta de la información de los alumnos cargada en la base de datos del proyecto.
6. Interfaz de usuario para la generación de reportes predefinidos en base de la información que arroje el modelo.
7. Informes predefinidos sobre la probabilidad de desertores en la carrera de Ingeniería Informática.
8. Documentación final del proyecto de tesis.

Criterios de aceptación del producto

1. El sistema contempla las funcionalidades que se han definido dentro de los requerimientos del proyecto. Se cumplen los requerimientos funcionales y no funcionales.
2. El resultado del modelo elegido nos muestra los posibles desertores indicando los factores más predominantes.
3. La interfaz del producto final es de fácil uso para el usuario.
4. La información final del documento de tesis es de fácil entendimiento para el usuario.

Estructura de Desglose del Trabajo y Entregables (EDT)

En la ilustración 11, podemos observar todos los entregables definidos en este trabajo.

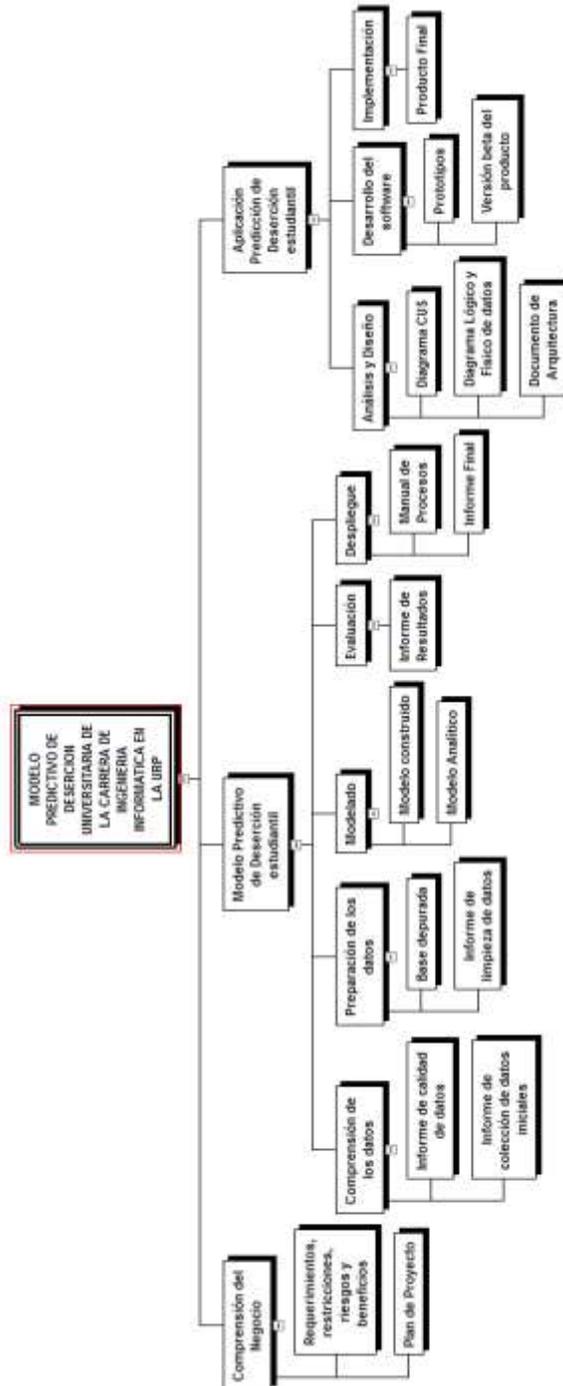


Ilustración 11: EDT del proyecto

La Estructura de Desglose del Trabajo y Entregables (EDT), está conformada por los entregables correspondientes a las dos metodologías usadas en la investigación, CRISP-DM para el Modelo Predictivo y RUP para el desarrollo de la aplicación y para ambos se comparten los entregables que recogen la definición de los requerimientos y plan de proyecto.

En la tabla 5, se describen los paquetes que conforman la EDT.

Tabla 5: Descripción de entregables de la EDT

Paquete	Descripción	Referencia
Requerimientos , restricciones, riesgos y beneficios	Requerimientos funcionales y no funcionales. Alcance, restricciones y riesgos del proyecto y producto. Beneficios.	Requerimientos del Producto/Software Importancia del Proyecto Restricciones del Proyecto Supuestos del Proyecto
Plan de Proyecto	Cronograma de actividades	Ver Anexo 4.
Informe de colección de datos iniciales	Descripción de datos iniciales	Ver Anexo 5.
Informe de calidad de datos	Descripción de la calidad de los datos recogidos inicialmente.	Ver Anexo 6.
Informe de limpieza de datos	Descripción de la limpieza a realizarse en los datos recogidos inicialmente.	Ver Anexo 8.
Base depurada	Base de datos depurada	Base de datos en SQL

	y limpia, apta para ser cargada y explotada.	Server
Modelo construido	Modelo construido con la base de datos depurada	Proyecto Solución en Análisis Services
Modelo Analítico	Modelo resultante.	Proyecto Solución en Analysis Services
Informe de Resultados	Resumen de resultados obtenidos luego de la minería de datos.	Capitulo Resultados
Informe Final	Informe Final, Tesis.	Tesis
Diagrama CUS	Diagrama general de los casos de uso del sistema	Diseño de la Solución - UML
Diagrama Lógico y Físico de datos	Modelo lógico y físico de datos	Diagrama de Clases Modelo Conceptual - UML
Documento de Arquitectura	Diagrama de despliegue y diagrama de arquitectura de la aplicación	Arquitectura del Producto/Software
Prototipos	Prototipos de la aplicación	Diseño de la Solución
Versión beta del producto	Primera Versión de la aplicación	Aplicación Beta
Producto Final	Versión final de la aplicación	Aplicación Final

Exclusiones del proyecto

1. El proyecto no brinda recomendaciones sobre las acciones que se deberán tomar frente a los casos de deserción que muestre el modelo.
2. El proyecto no utiliza otras herramientas de BI que no sean SQL Server Data Tools (Analysis Services e Integration Services)
3. El proyecto no contempla la instalación de los requerimientos técnicos necesarios para ejecutar el modelo.
4. El proyecto no brinda las licencias necesarias de las tecnologías usadas.
5. El proyecto no abarca los alumnos de las otras escuelas de la facultad de ingeniería.
6. El proyecto no contempla el costo de instalación de las herramientas tecnológicas usadas que no cuente la universidad.
7. Solo se capacitará una sola vez al personal autorizado que indique la escuela.
8. El proyecto no abarca la calibración futura del modelo.

Restricciones del proyecto

1. Costo: Solo se tuvo como presupuesto el costo estimado, no se incrementó posteriormente dicho presupuesto.
2. Tiempo: El proyecto tuvo un tiempo estimado de 5 meses de desarrollo y se limitó al uso de 2 recursos humanos.
3. Alcance: El proyecto se limitó a la información de alumnos de la escuela informática que brindó la universidad (OFICIC) por ende las variables del modelo de deserción se sacaron de la información brindada.
4. El proyecto se limitó al recurso técnico de las máquinas pertenecientes a las participantes del proyecto.
5. El proyecto se basó en las herramientas tecnológicas que cuenta la universidad.
6. Para el desarrollo del proyecto se utilizó la metodología CRISP-DM.
7. En el proyecto se usó la información de alumnos correspondiente a los últimos 4 años.

Supuestos del proyecto

1. Si uno de los participantes del proyecto se ausenta temporal o definitivamente, esto no deberá afectar las fechas comprometidas en el cronograma del proyecto.
2. Si una de las pc's donde se mostrará el producto falla, se tendrá otra pc de contingencia.
3. El personal de la universidad presentará una actitud colaboradora en todo momento del proyecto.
4. El personal de la universidad estará disponible para las fechas y horarios en que se realicen las entrevistas de levantamiento de información, dentro de los marcos temporales definidos para el proyecto, para lo cual las fechas y horas específicas de cada entrevista se fijarán en su momento y de común acuerdo entre el consultor y los miembros del equipo de proyecto.
5. La universidad cuenta con la información completa de los estudiantes para obtener las variables que se van analizar en el modelo.
6. Existe interés y compromiso de la universidad para acoger y potencializar las propuestas e iniciativas del proyecto.

2. Modelado de Negocio

En el presente capítulo se detalla el modelado del negocio, el diagrama de proceso y el diagrama de casos de uso del negocio.

Como proceso inicial (As Is), actualmente el negocio no contempla este modelo.

Como proceso futuro (To Be), se propone acceder a la información depurada de los estudiantes y ver resultados de tendencias de posibles desertores.

1. Caso de uso del negocio

En la ilustración 12, se muestra como proceso propuesto la escuela de informática realiza el caso de uso evaluación de deserción universitaria, que consiste en solicitar la información de los alumnos a OFICIC, luego procede a ejecutar el modelo y por último la generación de los reportes de los desertores.

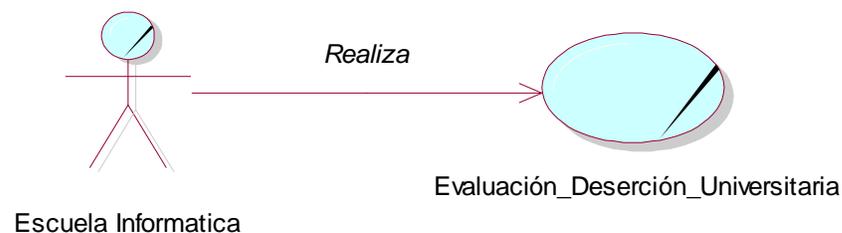


Ilustración 12: Diagrama de caso de uso del negocio, TO BE

2. Diagrama del proceso To Be, de evaluación de deserción universitaria

En la ilustración 13, se muestra el diagrama completo del proceso propuesto (To Be) y los actores que intervienen en cada tarea.

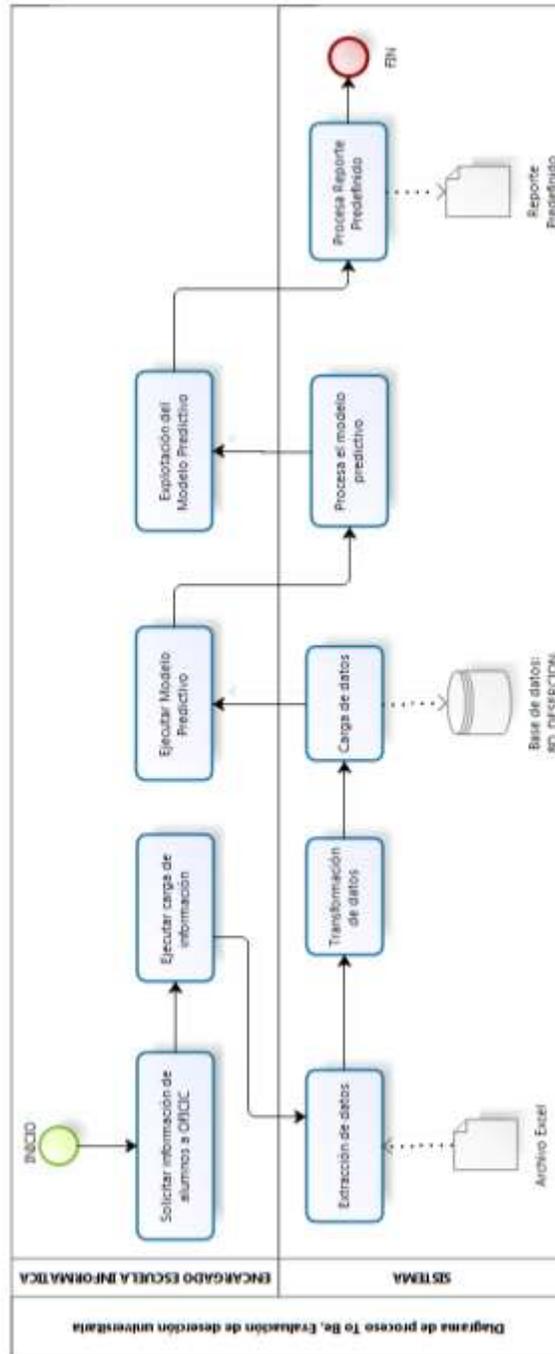


Ilustración 13: Diagrama del proceso propuesto (To Be)

3. Detalle del proceso To Be, de evaluación de deserción universitaria

A continuación en la tabla 6, se detalla cada paso del diagrama de proceso mostrado en la ilustración 13.

Tabla 6: Descripción del diagrama de proceso evaluación de deserción universitaria

Proceso	Descripción	Entrada	Salida
Solicitar información OFICIC	El encargado de la escuela solicita la información de los alumnos de la escuela de ingeniería informática a OFICIC	Pedido de información de alumnos a OFICIC.	Archivo Excel con la información de los alumnos.
Ejecutar Carga de información	El encargado de la escuela llena y guarda el archivo Excel con el formato establecido en la ruta que indicará el proyecto. Realiza la acción de	Archivo Excel con la información de los alumnos.	Archivo Excel con el formato establecido por el proyecto.

	ejecutar la carga de información.		
Extracción de datos	El sistema se encarga de realizar la lectura del archivo Excel y lo carga en una tabla temporal.	Archivo Excel con el formato establecido por el proyecto.	Tabla temporal en Integration Services (Tabla Stage).
Transformación de datos	El sistema depura la información de la tabla temporal Stage y realiza lógicas de algunos campos para guardarlos finalmente en las tablas del proyecto.	Tabla temporal en Integration Services(Tabla Stage).	Tablas del proyecto con información depurada y datos calculados de los alumnos.
Carga de datos	El sistema carga la información transformada de los alumnos	Tablas del proyecto con información depurada y datos calculados de los	Información de los alumnos cargada en las tablas de la Base de datos

	a las tablas del proyecto.	alumnos.	BD_DESERCION
Ejecutar modelo predictivo	El encargado de la escuela realiza la ejecución del modelo predictivo	Información de los alumnos cargada en las tablas de la Base de datos BD_DESERCION	Información de los alumnos cargada en las tablas de la Base de datos BD_DESERCION
Procesa el modelo predictivo	El sistema internamente en la herramienta Analysis Services procesa el modelo predictivo.	Información de los alumnos cargada en las tablas de la Base de datos BD_DESERCION	Información de los alumnos con posibilidad de desertar.
Explotación del modelo	El encargado de la escuela realiza la ejecución de explotación del modelo predictivo	Información de los alumnos con posibilidad de desertar que arroja el modelo predictivo.	El encargado de la escuela visualiza la información que arroja el modelo predictivo.
Procesa Reporte	El sistema internamente	Información de los alumnos con	Reportes predefinidos con

Predefinido	arma los reportes con la información que arroja el modelo predictivo	posibilidad de desertar que arroja el modelo predictivo.	la información del alumno desertor.
-------------	--	--	-------------------------------------

3. Requerimientos del Producto/Software

En este capítulo se verán todos los requerimientos funcionales y no funcionales de la aplicación.

3.1 Requerimientos funcionales

RF1: Extracción de información de alumnos

RF2: Limpiar información de la base de datos

RF3: Transformar información de los alumnos

RF4: Cargar información de los alumnos a la base de datos

RF5: Validar carga de información

RF6: Notificar ejecución del proceso ETL

RF7: Consultar información cargada de alumnos.

RF8: Procesar modelo predictivo de desertores

RF9: Consultar información del modelo de deserción

RF10: Generar reportes estadísticos de desertores

3.2 Matriz de trazabilidad

A continuación en la ilustración 14, se muestra los requerimientos funcionales vs los casos de uso del sistema.

Requerimientos funcionales / Casos de Uso del Sistema	CUS 1: Transformar y cargar datos	CUS2: Consultar información cargada	CUS3: Ejecutar Modelo	CUS4: Consultar Reportes
RF1: Extracción de información de alumnos				
RF2: Limpiar información de la base de datos				
RF3: Transformar información de los alumnos				
RF4: Cargar información de los alumnos a la base de datos				
RF5: Validar carga de información				
RF6: Notificar ejecución del proceso ETL				
RF7: Consultar información cargada de alumnos.				
RF8: Procesar modelo predictivo de desertores				
RF9: Consultar información del modelo de deserción				
RF10: Generar reportes estadísticos de desertores				

Ilustración 14: Matriz de trazabilidad

3.3 Requerimientos No Funcionales

RNF1: Disponibilidad del Sistema

La disponibilidad del sistema será de 99% del tiempo al mes, en condiciones normales.

RNF2: Requerimientos de rendimiento

El aplicativo debe tener un tiempo máximo de respuesta menor a 1 minuto.

RNF3: Requerimientos de soporte

Es una aplicación de escritorio que se encontrara localizada en un servidor local de la Institución.

RNF4: Usabilidad

Interfaz agradable para el usuario, las pantallas contarán con un formato pre-establecido, el cual evitara confusiones al usuario y le permitirá trabajar de una manera eficaz y práctica.

El sistema contara con validaciones, de tal manera que si hay un parámetro mal ingresado, sabrá responder mediante un mensaje adecuado.

El usuario se encontrará informado en todo momento sobre posibles errores al momento de interactuar con el sistema.

RNF5: Confiabilidad

El sistema debe ser tolerante a fallas. En caso de Fallas de algún componente no debe haber pérdida de información.

RNF6: Requerimientos de implementación

El desarrollo del sistema debe ser efectuado con la interacción entre un lenguaje de programación y una base de datos, que permitan desarrollar un sistema eficiente.

4. Arquitectura del Producto/Software

1. Vista de Despliegue

A continuación en la ilustración 15, describe la configuración de redes de trabajo físicas en el software desplegado y ejecutado. Podemos ver los elementos que participan en la red del sistema:

1. Usuario PC: Este elemento accederá a la aplicación.
2. Servidor de Base de Datos: En este elemento se alojará la base de datos en SQL Server, también se encontrará el paquete de carga y el modelo predictivo.
3. Aplicación: Este componente permite la interacción del usuario con el servidor mediante la realización de los casos de uso.

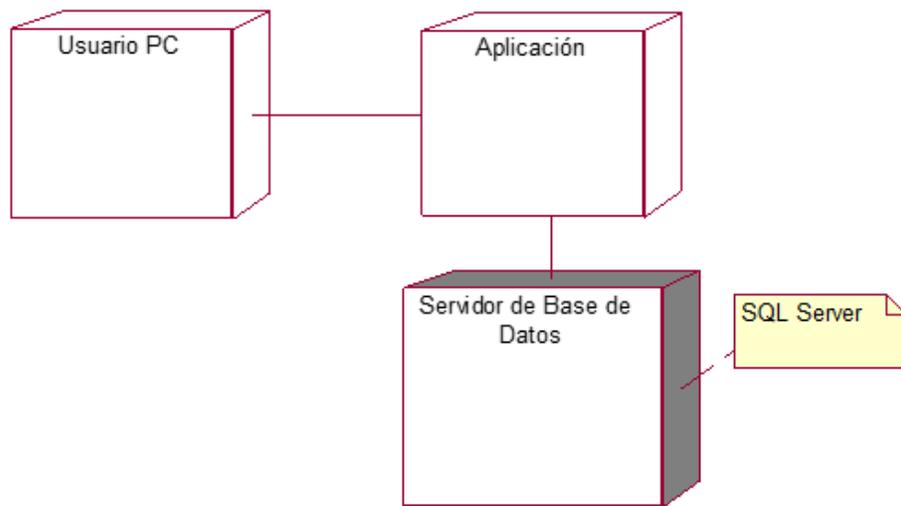


Ilustración 15 Diagrama de despliegue

2. Vista de implementación

A continuación en la ilustración 16, se muestra la descomposición del software entre capas y subsistemas.

Se observa la dependencia de la Aplicación a la Base de datos, así mismo la dependencia de la aplicación al gestor de base de datos SQL Server y a la plataforma .Net sobre la cual fue implementada.

La Aplicación se compone de los módulos de seguridad, carga de datos y consulta.

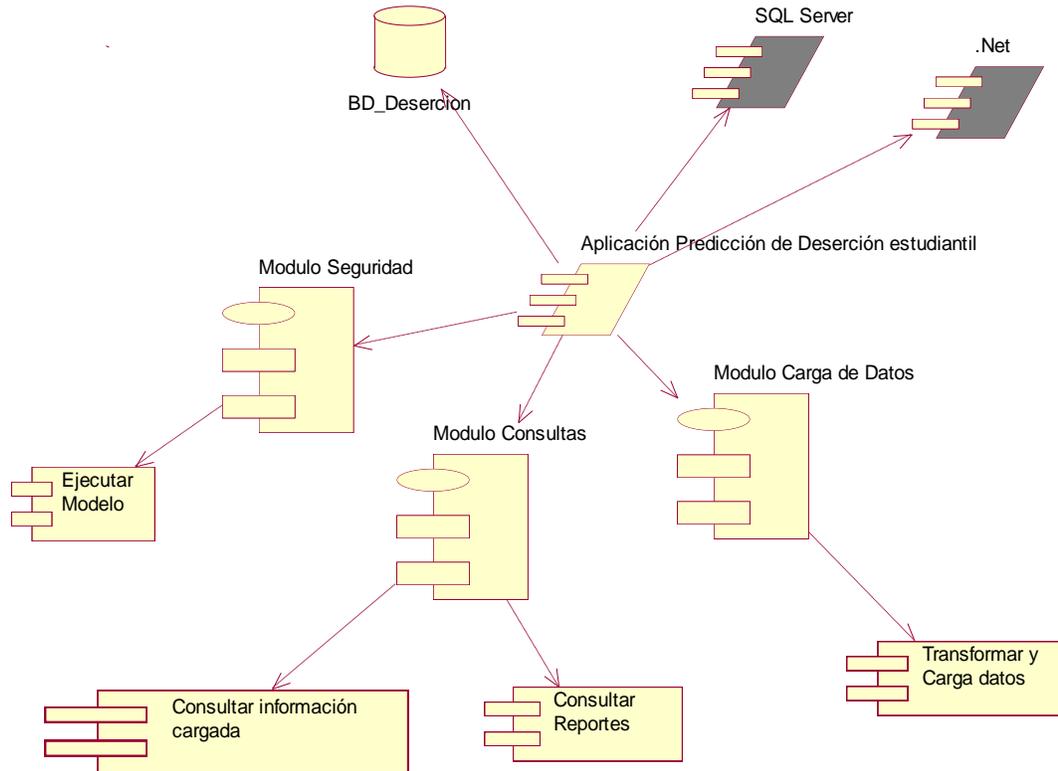


Ilustración 16: Diagrama de implementación

3. Vista General de la Arquitectura

En la ilustración 17, se observa la distribución en tres capas de arquitectura.

1. Capa de presentación

En esta capa se encuentra la aplicación que permitirá consultar la predicción de deserción de los alumnos de la escuela informática.

2. Capa Lógica de negocio

Esta capa provee la lógica del negocio, es decir la funcionalidad del sistema es decir la Carga y transformación de datos, la ejecución del modelo predictivo y las consultas.

3. Capa de acceso a datos

Esta capa provee la conexión a la base de datos requeridos por la capa de lógica de negocio.

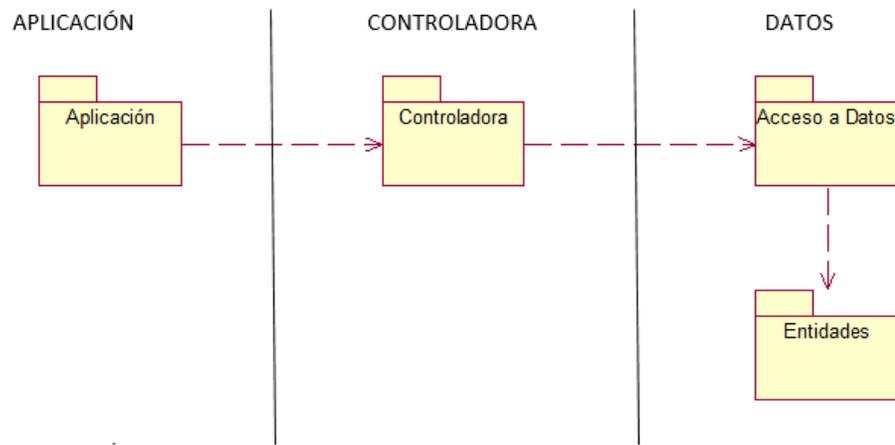


Ilustración 17: Diagrama de Capas

4. Vista lógica

En la ilustración18, se observa los paquetes del sistema, los cuales se describirán a continuación.

1. Carga de Datos

Este paquete contiene los casos de uso: Transformar y cargar datos.

2. Seguridad

Este paquete contiene los casos de uso que permiten el acceso al usuario al sistema: Iniciar Sesión y Cambiar Contraseña.

3. Consultas

Este paquete contiene los casos de uso: Consultar información cargada, Consultar Reportes y Ejecutar modelo

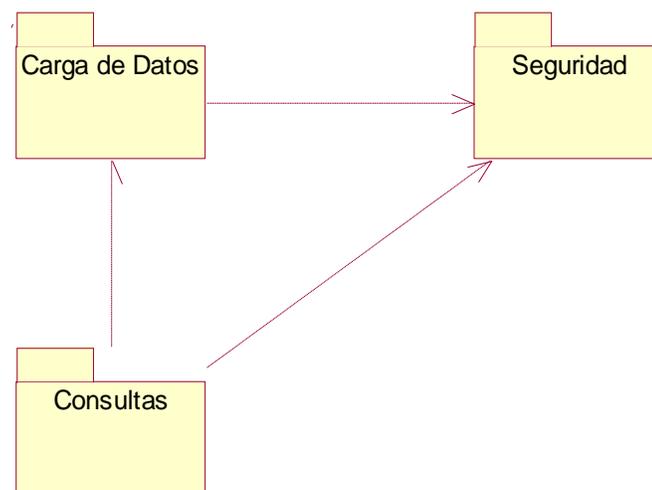


Ilustración 18: Diagrama de Paquetes del Sistema

5. Vista de datos

Esta vista describe la perspectiva de persistencia del sistema y como las clases se relacionan unas con otras.

3.1 Modelo Lógico

A continuación, en la ilustración 19 se muestra el diagrama Entidad – Relación del proyecto, clases, atributos y el tipo de datos.

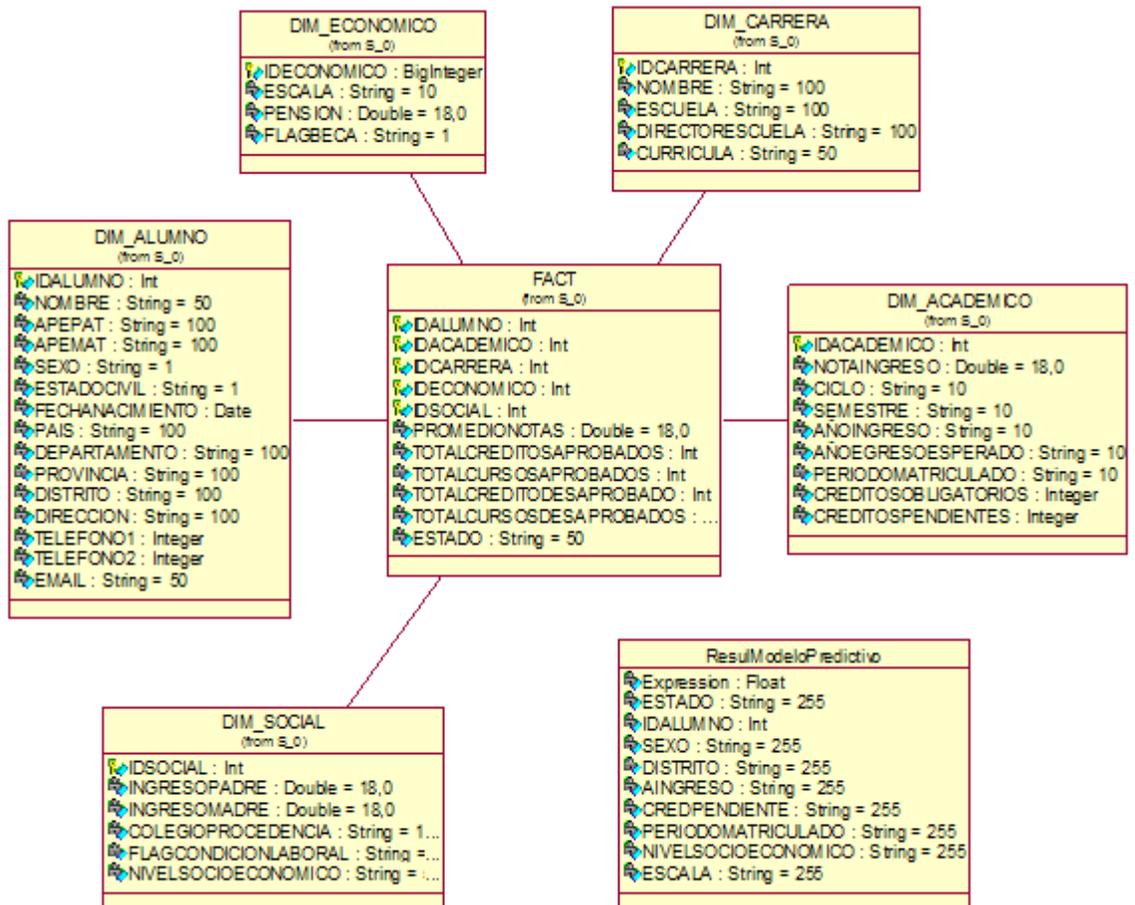


Ilustración 19: Diagrama Modelo Lógico (Entidad - Relación)

3.3 Modelo Físico

A continuación en la ilustración 20, se aprecia el modelo físico de las tablas de nuestra base de datos BD_DESERCION.

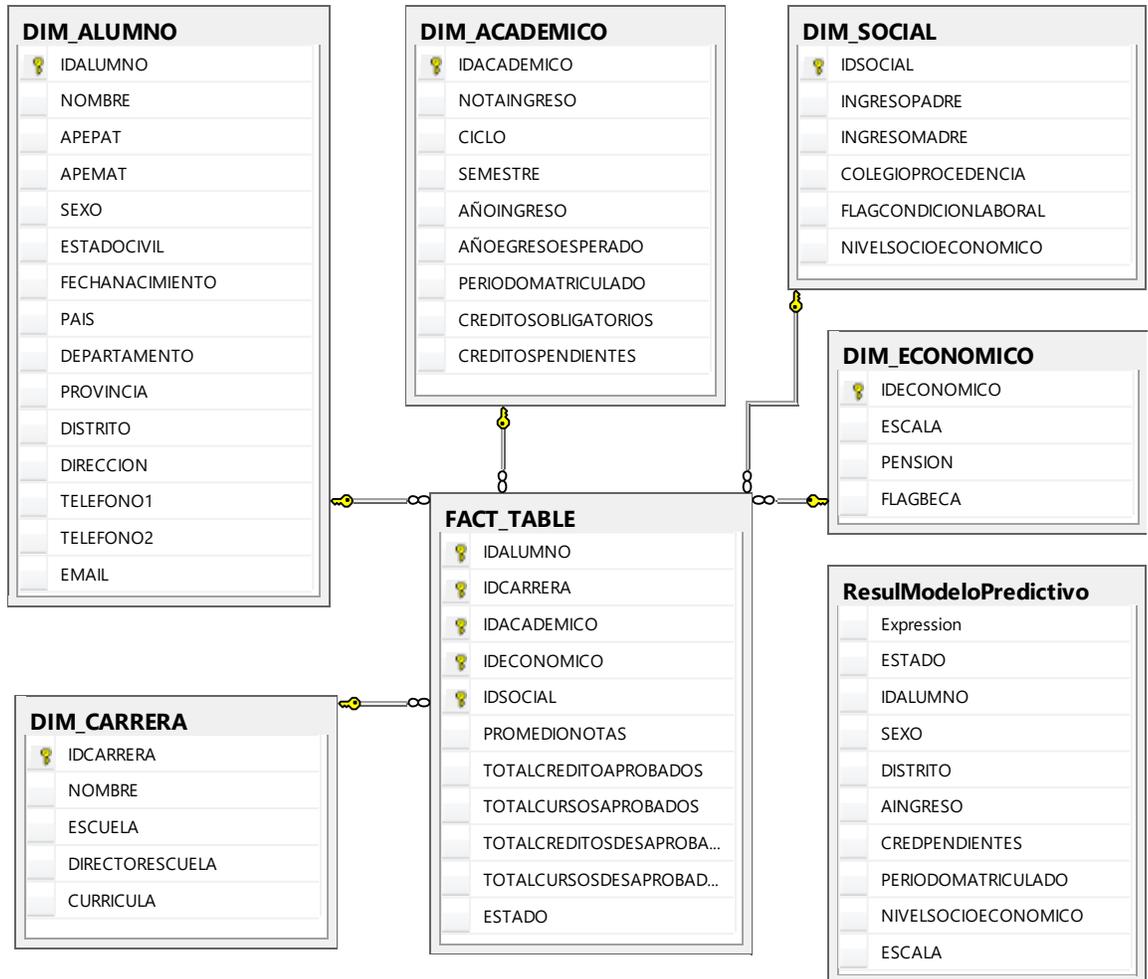


Ilustración 20: Diagrama Modelo Físico

6. Vista de casos de uso

Los diagramas de casos de uso sirven para facilitar la comunicación con los futuros usuarios del sistema, y resultan especialmente útiles para determinar las características necesarias que tendrá el sistema.

A continuación en la ilustración 21, se muestra el diagrama de casos de uso del proyecto, el cual describe las relaciones entre los casos de usos y el actor participante del proceso del modelo predictivo de deserción.

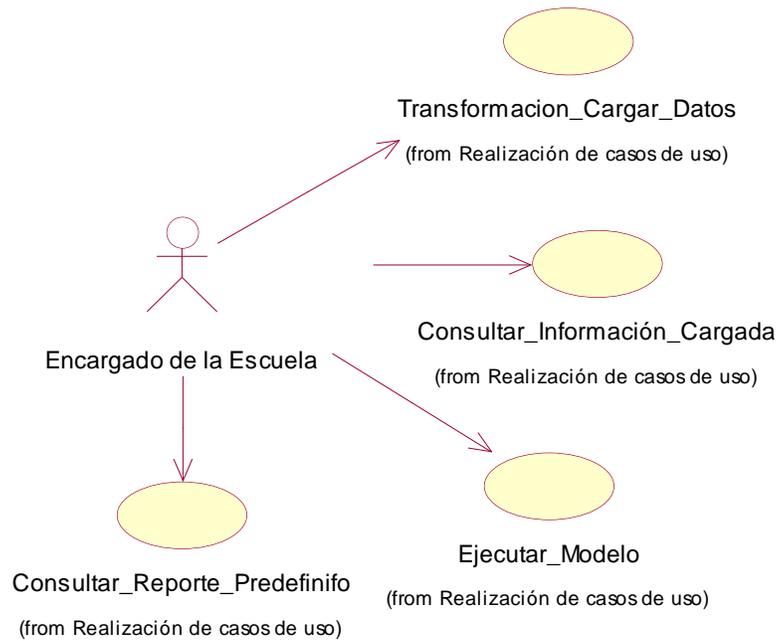


Ilustración 21: Vista de casos de uso del sistema

5. Diseño de la solución

A continuación se detallan los casos de uso del proyecto, el diagrama de clases, el diagrama de secuencia y los prototipos del producto.

5.1 Especificación de Casos de Uso del Sistema

1. CUS: Transformar y Cargar datos

En el siguiente caso de uso se inicia el proceso de extracción, transformación y carga de la información del alumnado de la escuela informática que alimentará el modelo estadístico mediante la cual se podrá predecir a los alumnos propensos a desertar.

El caso de uso empieza con el Encargado de la Escuela ingresando a la aplicación de escritorio en la cual presionará la opción “Ejecutar Carga de Información” que dará inicio al proceso ETL, cuyo primer paso es extraer la información del archivo Excel de una ruta predefinida de la PC local, luego se transforma la información del archivo, redefiniendo los tipos de variables y finalmente el proceso internamente carga como destino la información a la base de datos creada “BD_DESERCIÓN” la cual contiene la información lineal e histórica del estudiante que luego será explotada.

En la tabla 7, se muestran los actores que intervienen en este CUS con una breve descripción de la acción que realizan.

Actores Asociados

Tabla 7: Actores CUS, Transformar y Cargar datos

Actor	Tipo	Descripción
Encargado de la Escuela	Principal	Realiza la acción de carga de datos a la base de datos del proyecto
Sistema	Secundario	Ejecuta la acción enviada por el encargado de la Escuela.

Precondición

La persona encargada de la Escuela recibe por parte de OFICIC la información de alumnos de la escuela de ingeniería informática, luego procede a llenar el Excel entregado por el proyecto con un formato establecido de los campos que se van a cargar en la base de datos del proyecto. Culminado el llenado de la información se procede a guardar el archivo en la ruta definida en la PC local del usuario que el proyecto defina.

Flujo principal de eventos

Tabla 8: Especificación CUS, Transformar y Cargar datos

Paso	tipo	Descripción del paso
1	A	El encargado de la Escuela deberá ingresar a la aplicación de escritorio, al menú principal “Predicción de Deserción estudiantil – Escuela Ing. Informática”.
2	A	El encargado de la Escuela dará clic en la opción del menú, “Ejecutar Carga”
3	S	El sistema abrirá la pantalla de “Ejecución del proceso”.
4	A	El encargado de la Escuela dará clic en el botón

		“Ejecutar Carga de Información”
5	S	<p>El Sistema se conecta al servidor del SQL Server y ejecuta el trabajo “Ejecución_TransformarCargarDatos” creado en el Agente SQL Server.</p> <p>El trabajo “Ejecución_TransformarCargarDatos” llama al proceso ETL creado en la herramienta SQL Server Data Tools en la opción “Integration Services”.</p>
6	S	El sistema obtendrá la información del archivo origen Excel de la ruta establecida por el proyecto.
7	S	<p>El sistema dará inicio a la transformación:</p> <ol style="list-style-type: none"> 1. Se borra el contenido la tabla Stage. 2. Se borra el contenido de las tablas del datamart. 3. Se carga la información completa a la tabla Stage. 4. Se selecciona la información académica, se convierten los datos de ingreso en el formato de la tabla destino DIM_ACADEMICO 5. Se selecciona la información sobre datos personales de los alumnos. Se convierten los datos de ingreso en el formato de la tabla destino DIM_ALUMNO 6. Se selecciona la información sobre la escala y pensión de los alumnos. Se convierten los datos de ingreso en el formato de la tabla destino

		<p>DIM_ECONOMICA.</p> <p>7. Se selecciona la información de la carrera: Nombre de la escuela, código de la carrera, nombre del director de la escuela. Se convierten los datos de ingreso en el formato de la tabla destino DIM_CARRERA.</p> <p>8. Se selecciona la información sobre los padres de los alumnos. Se convierten los datos de ingreso en el formato de la tabla destino DIM_SOCIAL Se derivan los montos de ingreso a un solo campo por padre.</p> <p>9. Se selecciona la información del último periodo matriculado, créditos, estado de cursos de alumno. Se calcula la variable desertor. Se convierten los datos de ingreso en el formato de la tabla destino FACT_TABLE.</p>
8	S	<p>Los datos transformados se cargarán en la base de datos "BD_DESERCIÓN" establecida como destino.</p> <p>Finalizada la carga el sistema enviará automáticamente un mail a las direcciones establecidas previamente.</p> <p>El sistema mostrara en la pantalla un mensaje de carga satisfactoria del proceso.</p>
10	A	<p>El encardo de la escuela dará clic en el botón "Salir" para retornar al menú principal "Predicción de Deserción estudiantil – Escuela Ing. Informática".</p>

Flujo alternativo

Paso 4. Error en el paquete de carga

La información del log se mostrará en la pantalla, para que el encargado de solución. Ver anexo 1.

Paso	Tipo	Descripción del paso
1	S	El Sistema informa por pantalla y vía mail que la carga no fue exitosa.
2	A	El encargado de la escuela verifica la notificación del mail. Se deberá de ejecutar nuevamente el proceso haciendo clic en el botón "Ejecutar Carga de Información".
3	S	El sistema procesa nuevamente los datos mostrando un nuevo mensaje de confirmación y enviado vía mail que la carga fue exitosa.

Paso 5: No se Puede Conectar al Servidor SQL Server. Ver anexo 1.

Paso	Tipo	Descripción del paso
1	A	El encargado de escuela deberá revisar que estén activos los servicios de SQL.

Error Tipo: No se encuentra un elemento de conexión específico

Paso	Tipo	Descripción del paso
1	A	El encargado de escuela deberá verificar la integridad de la fuente de datos Excel y la base de datos destino.

Error Tipo: No se encuentra variable

Paso	Tipo	Descripción del paso
1	A	El encargado de escuela deberá verificar la integridad de la fuente de datos (Excel) y la base de datos destino, específicamente la existencia de las variables predefinidas.

Post condición

Se transformó y se cargó con éxito la información de los alumnos de la carrera de ingeniería informática.

Diagrama de secuencia

A continuación, en la ilustración 23 se muestra el intercambio de mensajes entre los objetos. Se observa como el actor principal, interactúa con la interfaz “Ejecutar Carga” y todo el proceso interno que realiza el sistema.

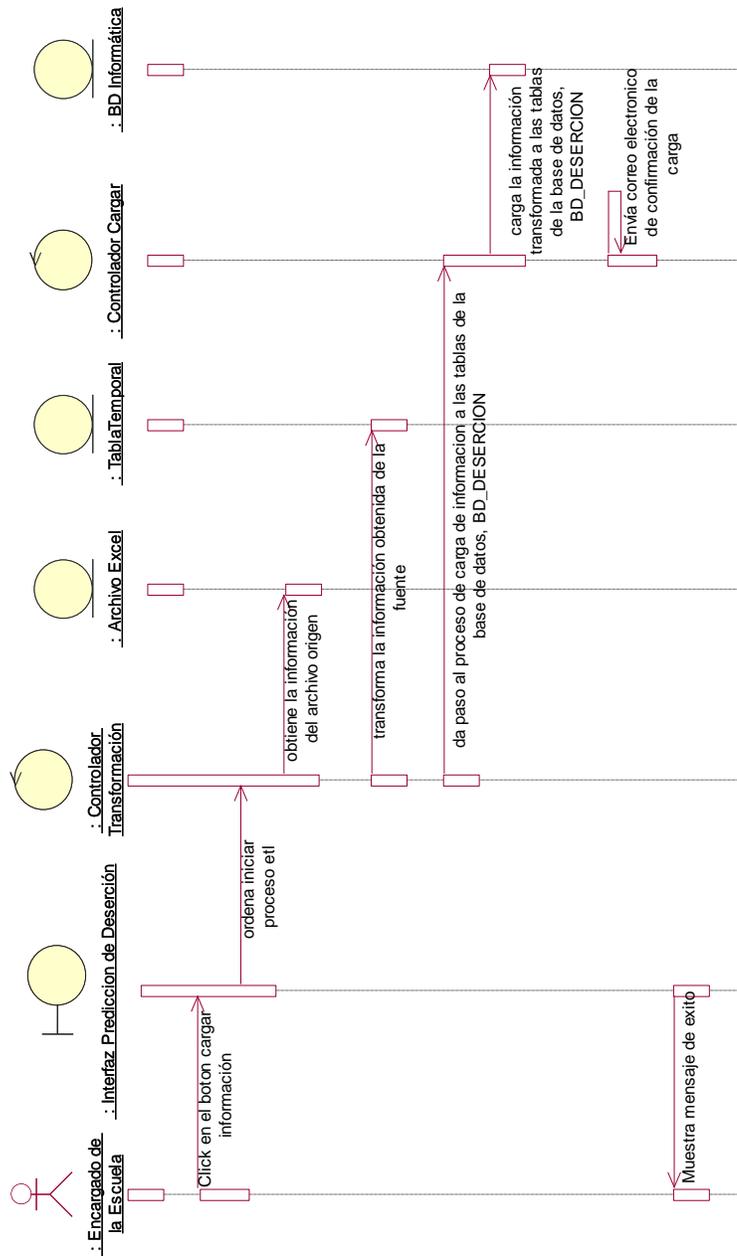


Ilustración 23: Diagrama de secuencia CUS, Transformar y Cargar Datos

Prototipo

1. Menú Principal: Interfaz “Predicción de Deserción estudiantil - Escuela Ing. Informática”

Para dar inicio al CUS Carga y Transformación de datos, seleccionamos la opción “Ejecutar Carga” como se muestra en la ilustración 24.



Ilustración 24: Menú Principal

Ejecutar Carga: Interfaz “Ejecución del proceso”

Al hacer clic en la opción “Ejecutar Carga” del menú principal “Predicción de Deserción estudiantil - Escuela Ing. Informática”, nos re direcciona a la interfaz “Ejecución del proceso” como se muestra en la ilustración 25.

Hacemos clic en el botón “Ejecutar Carga de Información” e internamente el sistema realiza la acción de extracción del archivo input (Excel con los datos de los alumnos), transformación de los datos con las lógicas establecidas y finalmente la carga de los datos hacia la base de datos del proyecto.



Ilustración 25: Interfaz “Ejecución del proceso”

Al terminar de ejecutarse el proceso ETL en el Integration Services, el sistema muestra el mensaje del resultado del trabajo ejecutado como se muestra en la ilustración 26.

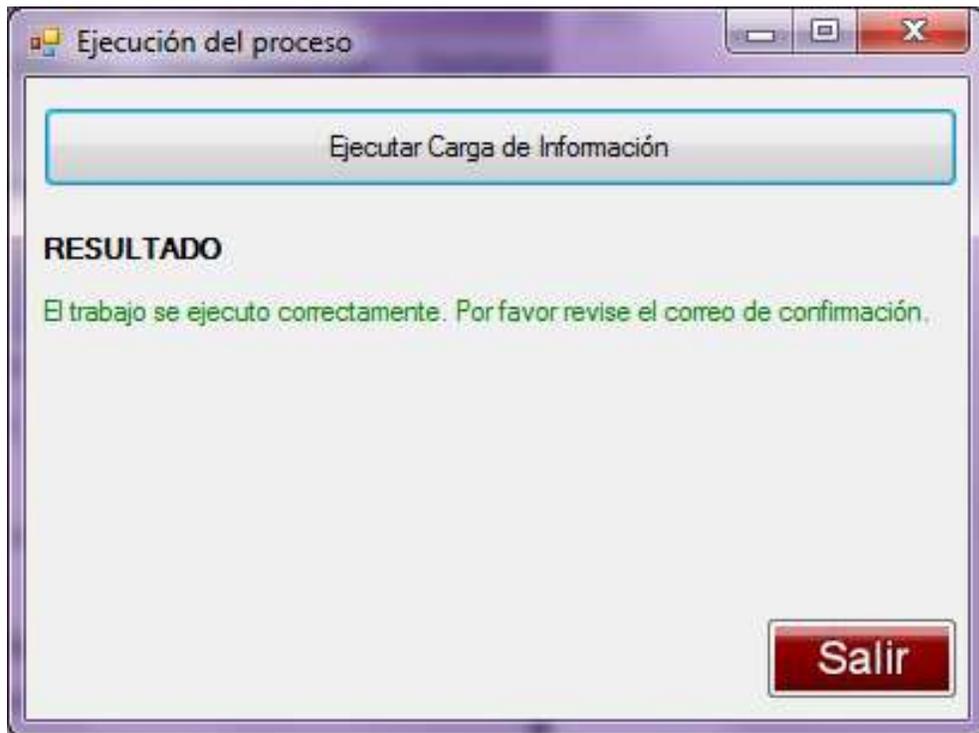


Ilustración 26: Ejecución correcta del ETL

2. CUS: Consultar información cargada

En el siguiente caso de uso se realiza el proceso de consulta de la información del alumnado de la escuela informática.

El caso de uso empieza con el Encargado de la Escuela ingresando a la aplicación de escritorio en la cual presionará la opción “Consultar Datos” y nos re direccionara a la interfaz “Consultar información cargada”, nos mostrara las opciones de consulta de la siguiente información: Carrera, Alumno, Socio-Económico y la información lineal del alumno

En la tabla 9, de muestran los actores que intervienen en este CUS con una breve descripción de la acción que realizan.

Actores Asociados

Tabla 9: Actores CUS, Consultar información cargada

Actor	Tipo	Descripción
Encargado de Escuela	Principal	Inicia el CUS: “Consultar Información Cargada”
Sistema	Secundario	Ejecuta la solicitud enviada por el Encargado de la Escuela

Precondición

Se ejecutó con éxito la carga de información a la base de datos “BD_DESERCION” del proyecto.

Flujo principal de eventos

Tabla 10: Especificación CUS, Consultar información cargada

Paso	tipo	Descripción del paso
1	A	El encargado de la Escuela deberá ingresar a la aplicación de escritorio, al menú principal “Predicción de Deserción estudiantil – Escuela Ing. Informática”.
2	A	El encargado de la Escuela dará clic en la opción del menú, “Consultar Datos”
3	S	El sistema abrirá la pantalla de “Consultar información cargada”.
4	A	El encargado de la Escuela dará clic en la opción que desea consultar.
5	S	El sistema mostrara la información seleccionada.
6	A	El encardo de la escuela dará clic en el botón “Salir” para retornar al menú principal “Predicción de Deserción estudiantil – Escuela Ing. Informática”.

Flujo alternativo

Paso 4: No se Puede Conectar al Servidor SQL Server

Paso	Tipo	Descripción del paso
1	A	El encargado de escuela deberá revisar que estén activos los servicios de SQL.

Post Condición

Se realiza con éxito la consulta de la información de los alumnos de la carrera de ingeniería informática.

Diagrama de clases CUS

En la ilustración 27 se muestra los elementos del diagrama de clases CUS Consultar Información Cargada (Actor, Interfaz, Controladora y Entidad).

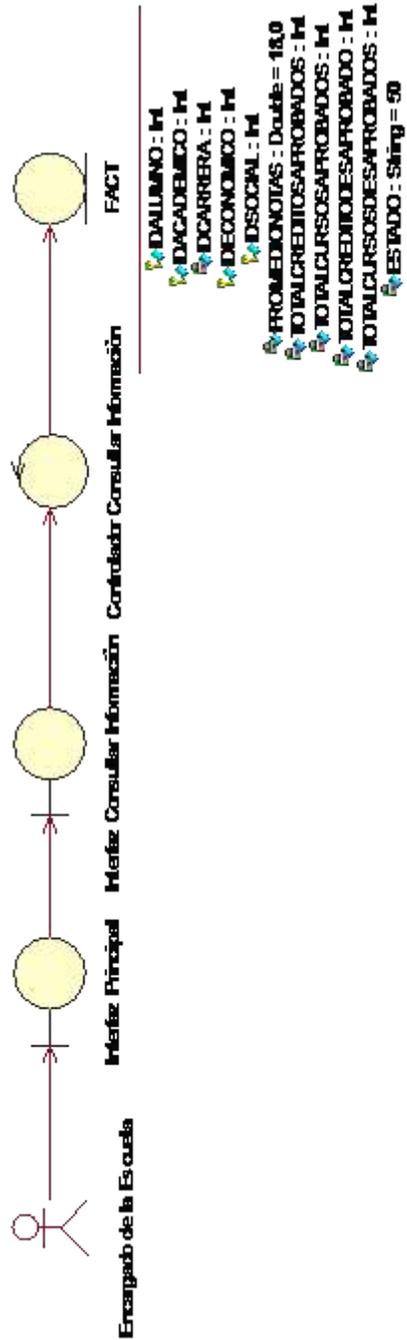


Ilustración 27: Diagrama de Clases CUS, Consultar Información Cargada

Diagrama de secuencia

A continuación, en la ilustración 28 se muestra el intercambio de mensajes entre los objetos. Se observa como el actor principal, interactúa con la interfaz "Consultar información cargada" y todo el proceso interno que realiza el sistema.

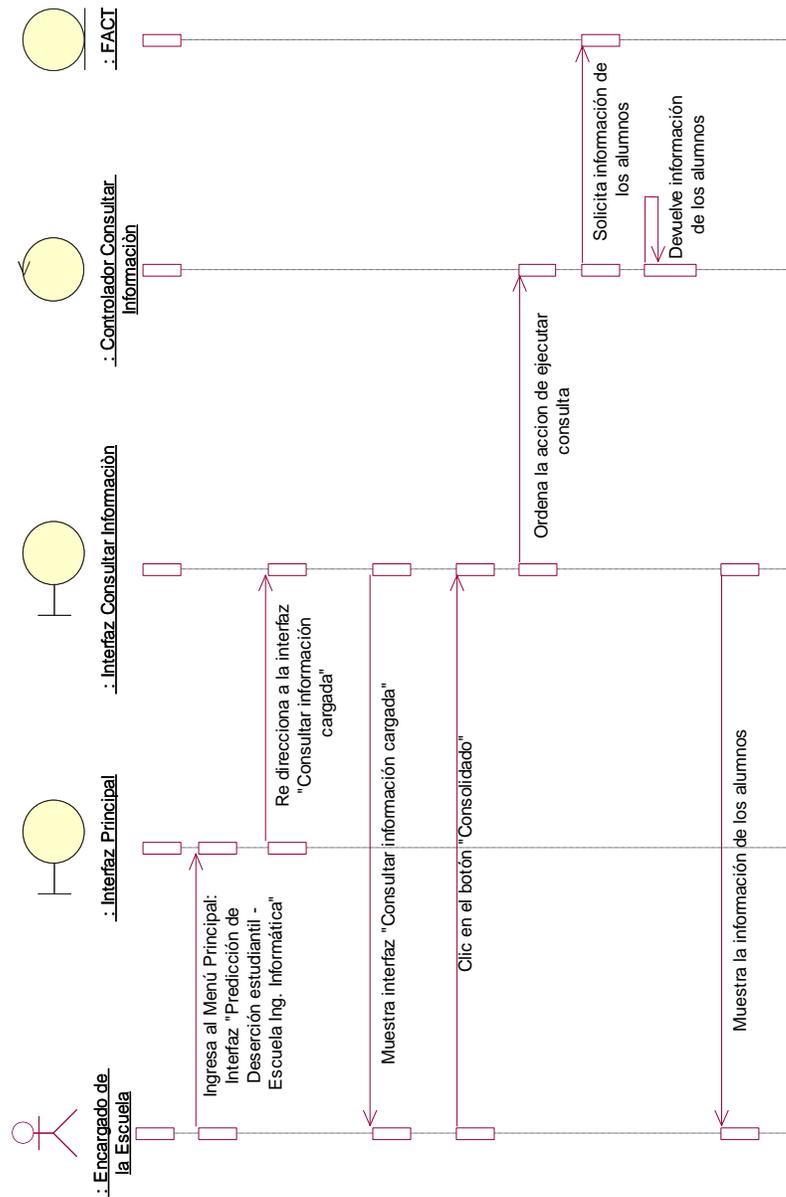


Ilustración 28: Diagrama de secuencia CUS, Consultar Información Cargada

Prototipo

1. Menú Principal: Interfaz “Predicción de Deserción estudiantil - Escuela Ing. Informática”

Para dar inicio al CUS Consultar información Cargada, seleccionamos la opción “Consultar Datos” como se muestra en la ilustración 24.

2. Consultar Datos: Interfaz “Consultar información cargada”

Al hacer clic en la opción “Consultar Carga” del menú principal “Predicción de Deserción estudiantil - Escuela Ing. Informática”, nos re direcciona a la interfaz “Consultar información cargada” como se muestra en la ilustración 29.

La interfaz nos muestra 4 opciones de consulta sobre la información de los alumnos carga y depurada.

1. Carrera: Nos muestra la información de la carrera del alumno.
2. Alumno: Nos muestra los datos personales del alumno.
3. Socio – Económico: Tenemos dos opciones a elegir de consulta, tanto información social del alumno como información económica del alumno.
4. Consolidado: Nos muestra la información consolidada de la historia del alumno en la carrera.

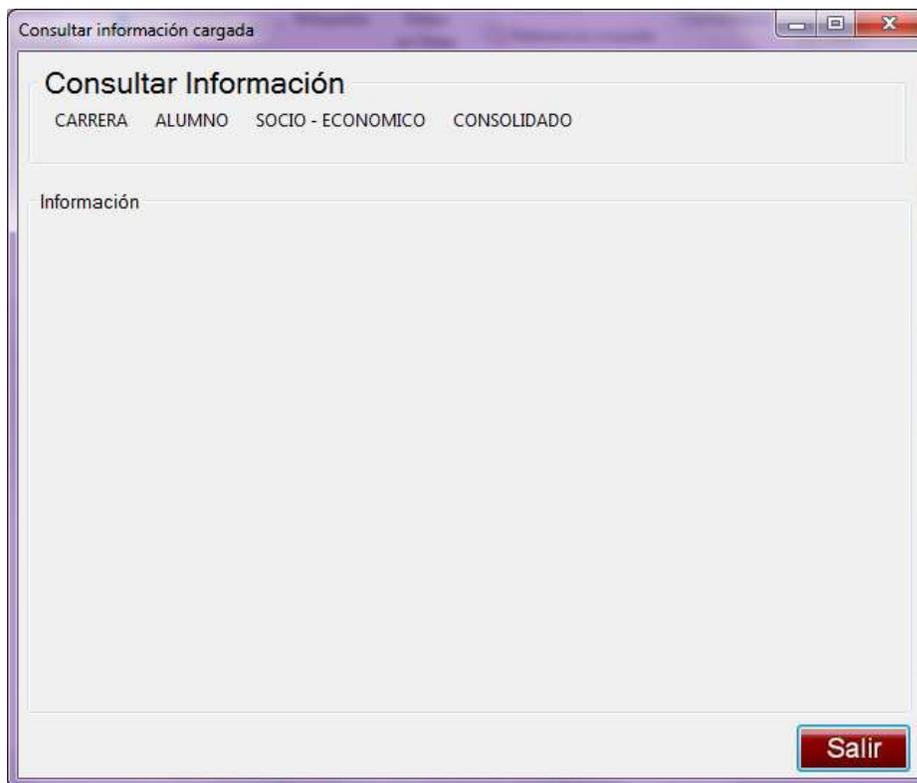


Ilustración 29: Interfaz "Consultar información cargada"

Hacemos clic en el botón "Consolidado" e internamente el sistema realiza la acción de consulta a la base de datos "BD_DESERCION" y nos muestra la información solicitada como se muestra en la ilustración 30.

Consultar información cargada

Consultar Información

CARRERA ALUMNO SOCIO - ECONOMICO **CONSOLIDADO**

Información consolidada del alumno

	Cod. Alumno	Carrera	Promedio de notas	Total creditos aprobados	Total cursos aprobados	Total desap
▶	99710496	66	3	36	0	13
	99710591	66	4	193	0	3
	99710729	66	12	201	20	4
	99711002	66	11	200	1	0
	99711211	66	12	221	39	9
	99711484	66	1	56	0	5
	99714268	66	99	177	0	0
	99724491	66	13	189	8	19
	99724499	66	12	203	1	0
	99806236	66	2	197	0	5
	99818277	66	13	201	1	0
	99835249	66	9	66	2	14

Salir

Ilustración 30: Información consolidado del alumno

3. CUS: Ejecutar Modelo

En el siguiente caso de uso se inicia el proceso de ejecución del modelo predictivo, el cual se podrá predecir a los alumnos propensos a desertar.

El caso de uso empieza con el Encargado de la Escuela ingresando a la aplicación de escritorio en la cual presionará la opción “Ejecutar Modelo Predictivo” que dará inicio al proceso de ejecución del modelo predictivo en la herramienta Analysis Service, cuyo primer paso es extraer la información de la base de datos creada “BD_DESERCIÓN”, la cual contiene la información lineal e histórica de los estudiantes de informática de la Universidad Ricardo Palma. Una vez obtenida la información se usarán las técnicas de minería de datos que provee la herramienta Analysis Service, entre las técnicas que se van a comparar son: Arboles de decisión y Clusters.

Una vez finalizada la ejecución del modelo, se podrá consultar los resultados del modelo.

En la tabla 11, de muestran los actores que intervienen en este CUS con una breve descripción de la acción que realizan.

Actores Asociados

Tabla 11: Actores CUS, Ejecutar Modelo

Actor	Tipo	Descripción
Encargado de la escuela	Principal	Ejecuta el Modelo de predicción
Sistema	Secundario	Realiza el procesamiento del modelo predictivo.

		Ejecuta la consulta de los resultados del modelo predictivo.
--	--	--

Pre condición

El proceso de carga de datos deberá de ser exitoso.

Flujo principal de eventos

Tabla 12: Especificación CUS, Ejecutar Modelo

Paso	tipo	Descripción del paso
1	A	El encargado de la Escuela deberá ingresar a la aplicación de escritorio, al menú principal “Predicción de Deserción estudiantil – Escuela Ing. Informática”. El encargado de la Escuela dará clic en la opción del menú, “Modelo Predictivo”, el cual despliega la opción “Ejecutar Modelo Predictivo”.
2	S	El sistema abrirá la pantalla de “Ejecución del Modelo Predictivo”.
3	A	El encargado de la Escuela dará clic en el botón “Ejecutar Modelo Predictivo”.
4	S	El Sistema se conecta al servidor del SQL Server y ejecuta el trabajo “Ejecucion_ModeloPredictivo” creado en el Agente SQL Server. El trabajo “Ejecucion_ModeloPredictivo” llama al proceso de minería de datos “Modelo_DesercionUniversitaria_URP”, creado en la

		<p>herramienta SQL Server Data Tools en la opción “Analysis Services”.</p> <p>Analysis Services realiza el procesamiento del modelo predictivo y guarda la información de los resultados en la tabla “ResulModeloPredictivo”.</p> <p>Finalizada la ejecución del modelo, el sistema enviará automáticamente un mail a las direcciones establecidas previamente.</p> <p>El sistema mostrara en la pantalla un mensaje de carga satisfactoria del proceso.</p>
5	A	El encardo de la escuela dará clic en el botón “Salir” para retornar al menú principal “Predicción de Deserción estudiantil – Escuela Ing. Informática”.
6	A	El encargado de la Escuela dará clic en la opción del menú, “Modelo Predictivo”, el cual despliega la opción “Consultar Resultado del Modelo”.
7	S	El sistema muestra la pantalla “Consulta de resultados del modelo”
8	A	El encargado de la Escuela dará clic en el botón “Consultar Resultado Modelo Predictivo”.
9	S	El sistema muestra la información que arrojo el modelo predictivo indicando que alumnos son los posibles Desertores.
10	A	El encardo de la escuela dará clic en el botón “Salir” para retornar al menú principal “Predicción de Deserción estudiantil – Escuela Ing. Informática”.

Flujos alternativos

Paso 4. Error en el paquete de ejecución del modelo

La información del log se mostrará en la pantalla, para que el encargado de solución. Ver anexo 1.

Paso	Tipo	Descripción del paso
1	S	El Sistema informa por pantalla y vía mail que la carga no fue exitosa.
2	A	El encargado de la escuela verifica la notificación del mail. Se deberá de ejecutar nuevamente el proceso haciendo clic en el botón "Ejecutar Modelo Predictivo".
3	S	El sistema procesa nuevamente los datos mostrando un nuevo mensaje de confirmación y enviado vía mail que la ejecución del modelo predictivo fue exitosa.

Paso 4: No se Puede Conectar al Servidor SQL Server. Ver anexo 1.

Paso	Tipo	Descripción del paso
1	A	El encargado de escuela deberá revisar que estén activos los servicios de SQL.

Post Condición

Se realizó la ejecución del modelo de predicción con éxito.

Diagrama de clases CUS

En la ilustración 31 se muestra los elementos del diagrama de clases CUS Ejecutar Modelo (Actor, Interfaz, Controladora y Entidad).

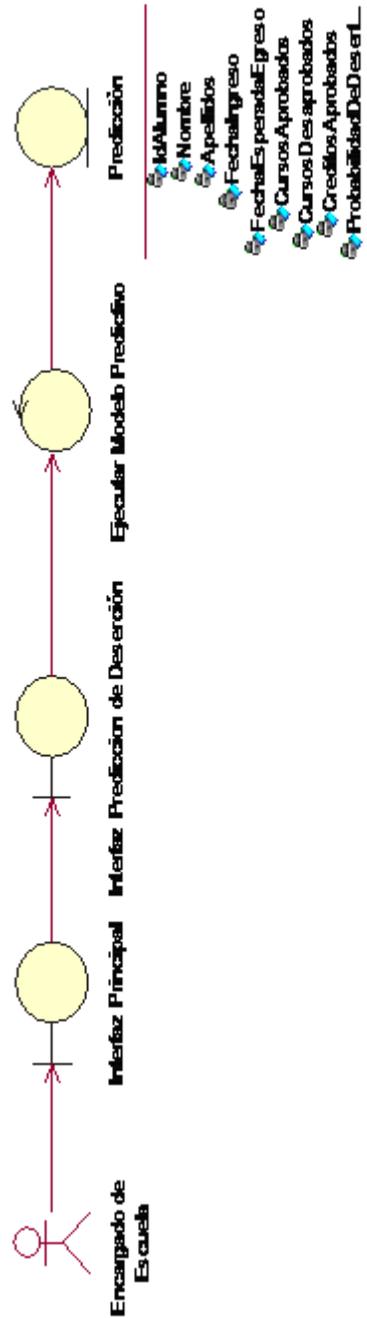


Ilustración 31: Diagrama de clases CUS, Ejecutar Modelo

Diagrama de secuencia

A continuación, en la ilustración 32 se muestra el intercambio de mensajes entre los objetos. Se observa como el actor principal, interactúa con la interfaz “Consulta de resultados del modelo” y todo el proceso interno que realiza el sistema.

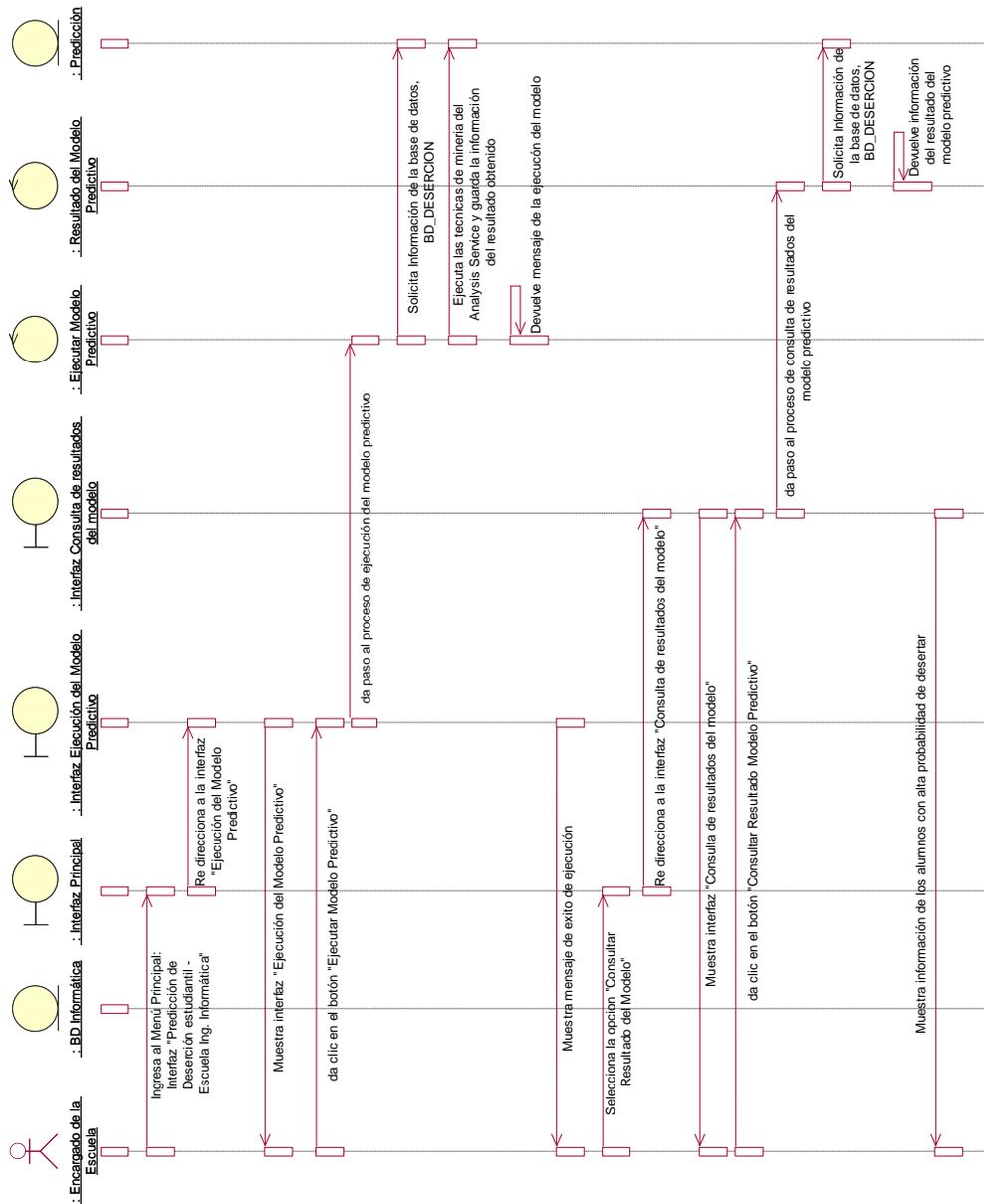


Ilustración 32: Diagrama de secuencia CUS, Ejecutar Modelo

Prototipos

1. Menú Principal: Interfaz “Predicción de Deserción estudiantil - Escuela Ing. Informática”

Para realizar la ejecución del modelo predictivo y la consulta de los resultados obtenidos, seleccionamos la opción “Modelo Predictivo” como se muestra en la ilustración 24.

2. Ejecutar Modelo Predictivo: Interfaz “Ejecución del Modelo Predictivo”

Al hacer clic en la opción “Ejecutar Modelo Predictivo” del menú principal “Predicción de Deserción estudiantil - Escuela Ing. Informática”, nos re direcciona a la interfaz “Ejecución del Modelo Predictivo” como se muestra en la ilustración 33.

Hacemos clic en el botón “Ejecutar Modelo Predictivo” e internamente el sistema realiza la acción de minería de datos, consulta la información de los alumnos de la base de datos BD_DESERCION y utiliza las técnicas de minería de datos: arboles de decisión y clústeres.

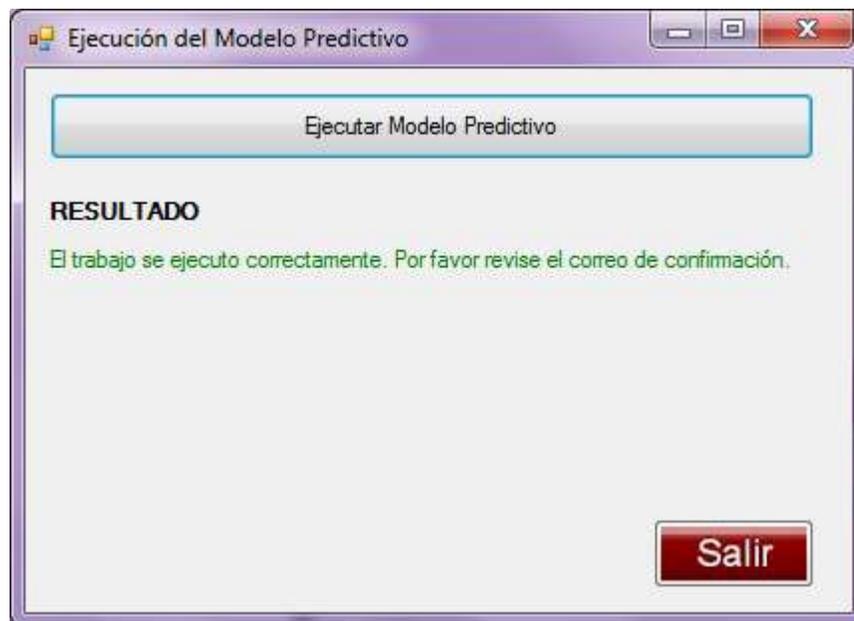


Ilustración 33: Interfaz "Ejecución del Modelo Predictivo"

3. Consultar Resultado del Modelo: Interfaz "Consulta de resultados del modelo"

Al hacer clic en la opción "Consultar Resultado del Modelo" del menú principal "Predicción de Deserción estudiantil - Escuela Ing. Informática", nos re direcciona a la interfaz "Consulta de resultados del modelo" como se muestra en la ilustración 34.

Hacemos clic en el botón "Consultar Resultado Modelo Predictivo", el sistema nos muestra la información que el modelo predictivo arroja.

Consulta de resultados del modelo

Consultar Resultado Modelo Predictivo

RESULTADO

	Expression	Estado	Cod. Alumno	Sexo	Distrito	Año de Ingreso
▶	0.5277777777777779	Desertor	101011841	M	OTROS	20101
	0.5277777777777779	Desertor	101011842	M	Lima Sur	20101
	0.5277777777777779	Desertor	101011843	M	OTROS	20101
	0.5277777777777779	Desertor	101011849	M	OTROS	20101
	0.5277777777777779	Desertor	101011853	M	Lima Sur	20101
	0.5277777777777779	Desertor	101011854	M	Lima Oeste	20101
	0.5277777777777779	Desertor	101011856	M	OTROS	20101
	0.5277777777777779	Desertor	101011996	M	Lima Oeste	20101
	0.90217391304347827	Desertor	101012105	M	Lima Sur	20101
	0.5277777777777779	Desertor	101012758	M	Lima Este	20101
	0.5277777777777779	Desertor	101012767	M	OTROS	20101
	0.5277777777777779	Desertor	101012768	M	OTROS	20101

Salir

Ilustración 34: Consulta de resultados del modelo

4. CUS: Consultar Reporte Predefinido

El encargado de la escuela de ingeniería informática podrá acceder a la opción generar reporte a través de una interfaz amigable e intuitiva. Una vez generado el modelo predictivo, el encargado podrá realizar la consulta de la historia cargada en cualquier momento.

En la tabla 13, se muestran los actores que intervienen en este CUS con una breve descripción de la acción que realizan.

Actores Asociados

Tabla 13: Actores CUS, Consultar reporte predefinido

Actor	Tipo	Descripción
Encargado de escuela	Principal	Realiza la acción generar reporte
Sistema	Secundario	Procesa la acción solicitada por el encargado de la escuela

Pre condición

Se deberá haber realizado con éxito el CUS Ejecutar Modelo.

Flujo principal de eventos

Tabla 14: Especificación CUS, Consultar reporte predefinido

Paso	Tipo	Descripción del paso
1	A	El encargado de la Escuela deberá ingresar a la aplicación de escritorio, al menú principal "Predicción de Deserción estudiantil – Escuela Ing. Informática".

2	A	El encargado de la Escuela dará clic en la opción del menú, "Reporte"
3	S	El sistema muestra las siguientes opciones de reporte a generar: Consolidado, Detallado.
4	A	El encargado de la Escuela dará clic en el botón sobre una de las opciones.
5	S	El sistema obtiene la información transformada de la base de datos y que arrojo el modelo de deserción y arma el reporte. El sistema muestra el reporte generado

Flujo alternativo

Paso 4. Error en la generación del reporte

Paso	Tipo	Descripción del paso
1	S	Sistema muestra un mensaje de error.
2	A	El encargado cierra la interfaz y vuelve a ingresar.

Post condición

Se realizó la consulta con éxito.

Diagrama de clases CUS

En la ilustración 35 se muestra los elementos del diagrama de clases CUS Consultar Reporte Predefinido (Actor, Interfaz, Controladora y Entidad).

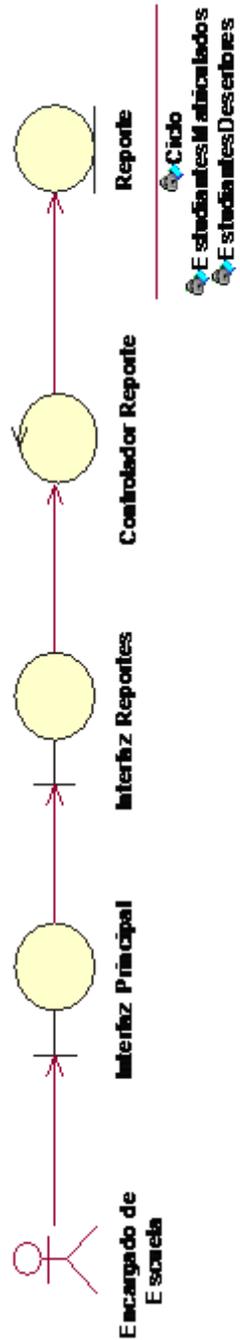


Ilustración 35: Diagrama de Clases CUS, Consultar Reporte Predefinido

Diagrama de secuencia

A continuación, en la ilustración 36 se muestra el intercambio de mensajes entre los objetos. Se observa como el actor principal, interactúa con la interfaz “Reporte” y todo el proceso interno que realiza el sistema.

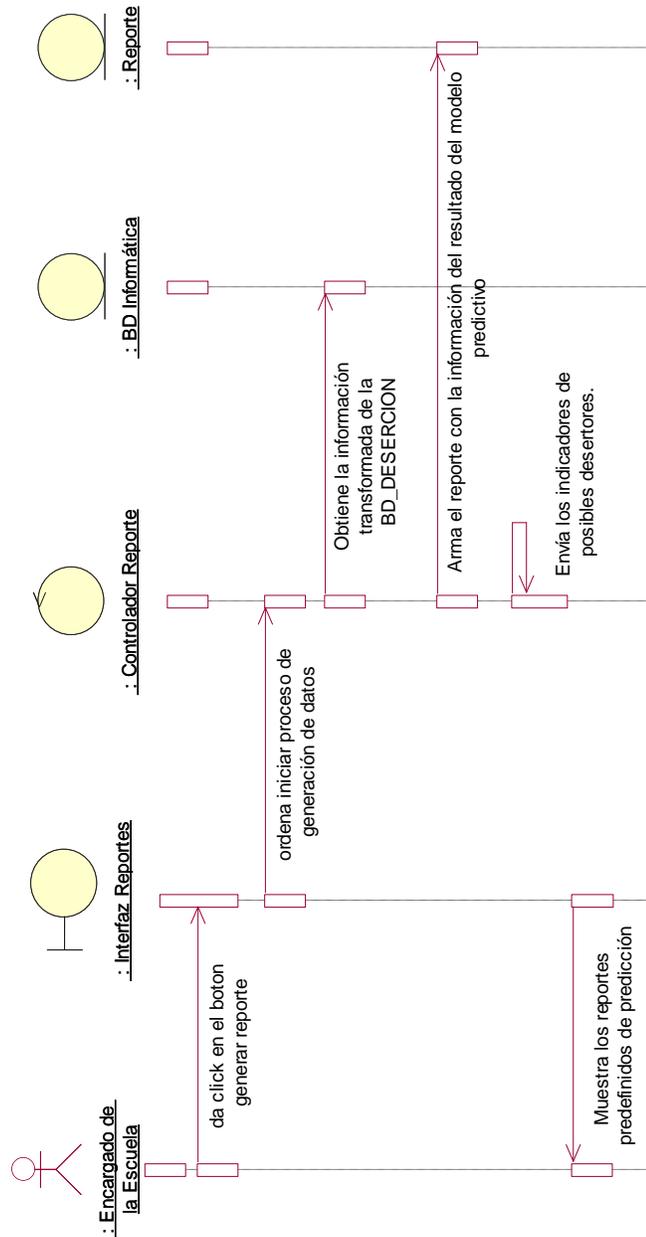


Ilustración 36: Diagrama de Secuencia CUS, Consultar Reporte Predefinido

Prototipos

1. Menú Principal: Interfaz “Predicción de Deserción estudiantil - Escuela Ing. Informática”

Para dar inicio al CUS Consultar Reporte Predefinido, seleccionamos la opción “Reportes” como se muestra en la ilustración 24.

2. Reportes

Al hacer clic en la opción “Ejecutar Modelo” del menú principal “Predicción de Deserción estudiantil - Escuela Ing. Informática”, nos muestra las opciones de consulta, reporte detallado y reporte consolidado. Seleccionamos las opciones y se muestran los reportes de la ilustración 37 y 38.

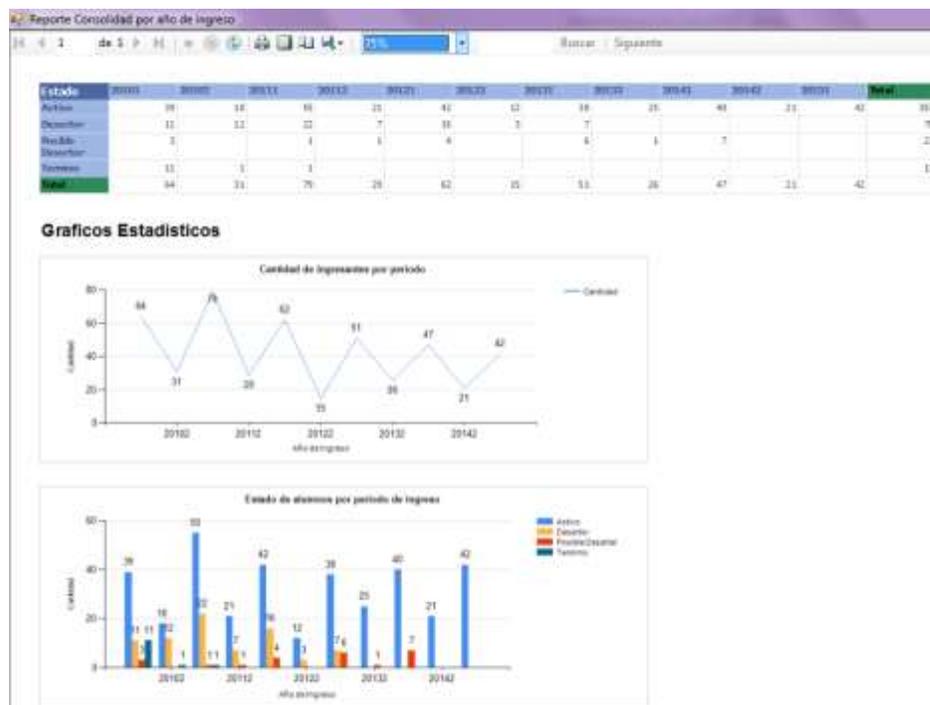


Ilustración 37: Reporte Consolidado por año de ingreso

Detalle

de 57

75%

Buscar | Siguiente



Código	Distrito	Sexo	Año de ingreso	Último periodo matriculado	Creditos Pendientes	Pensión S/	Estado	Indicador
101021442	Pueblo Libre	M	20102	20121	191	990	Desertor	●
101021445	Vila Maria del Trunfo	M	20102	20111	205	990	Desertor	●
101021446	Vila Maria del Trunfo	M	20102	20121	163	990	Desertor	●
101021448	Vila Maria del Trunfo	H	20102	20111	190	990	Desertor	●
101021449	Santiago de Surco	M	20102	20111	198	990	Desertor	●
101021450	Banasco	M	20102	20111	214	990	Desertor	●
101021451	Vila Maria del Trunfo	M	20102	20111	218	990	Desertor	●
101021452	Ala	H	20102	20111	176	990	Desertor	●
101021454	Celico	M	20102	20121	228	990	Desertor	●
101021892	San Juan de Lurigancho	H	20102	20121	158	990	Desertor	●
101021285	San Juan de Miraflores	H	20102	20112	203	990	Desertor	●
101021896	Carmel	M	20102	20142	188	990	Posible Desertor	●
101021065	Arcos	M	20102	20142	168	990	Posible Desertor	●
101111858	La Perla	M	20111	20151	61	990	Adiivo	●
101111857	Lima	M	20111	20151	50	990	Adiivo	●
101111866	Lima	H	20111	20151	165	990	Adiivo	●
101111870	Imperial	M	20111	20151	61	990	Adiivo	●
101113338	El Aguafino	M	20111	20151	48	990	Adiivo	●
101113338	Vila B Salvador	F	20111	20131	38	990	Adiivo	●
101113346	Vila B Salvador	M	20111	20151	98	990	Adiivo	●
101113342	Santiago de Surco	H	20111	20151	92	990	Adiivo	●

Ilustración 38: Reporte Detallado

6. Evaluación de Resultados

6.1 Objetivos

En referencia a los objetivos establecidos en esta investigación en el punto Propósito del Proyecto dentro del Capítulo de introducción, se cumplió el implementar un modelo predictivo con la información del alumnado de los últimos 4 años de la escuela informática usando la metodología CRISP-DM para los procesos de minería de datos y RUP para la aplicación desarrollada.

SQL Server Analysis Services proporciona una interfaz de diseño gráfica para crear consultas y un lenguaje de consulta denominado Extensiones de minería de datos (DMX) que resulta de gran utilidad para crear predicciones personalizadas, las cuales nos permitieron identificar a los alumnos con alta probabilidad de abandonar la carrera de ingeniería informática.

Los resultados del procesamiento del modelo mostraron las principales causas de la deserción en la escuela informática y también nos permitieron alimentar reportes donde se trabajaron indicadores.

6.2 Datos procesados

El registro de los datos input se realizó en la herramienta de BI Integration Services, el cual satisface los objetivos de carga del proceso ETL.

En cada tarea se transforma la data que brindó OFICIC, finalmente obtenemos data sin repitencia y con el formato correcto que se carga en la base de datos estructurada.

Se calcularon variables como el nivel socioeconómico a partir del distrito de residencia, la escala y el ingreso económico de los padres.

En la ilustración 39, podemos observar que el proceso de carga de datos en cada una de las tablas de nuestro proyecto, es realizado con éxito y se verifica con el check verde que se aprecia en la ilustración.

La ejecución del proceso inicia con la eliminación de los datos de la temporal que guarda la información del archivo Excel, luego continua con la eliminación de los datos de las tablas del proyecto.

Una vez eliminada la información, se cargan los nuevos valores del archivo Excel a la tabla temporal Stage, luego prosigue a cargar todas las tablas de proyecto. Durante el transcurso de carga se realiza la transformación de los datos y cálculos de algunas variables.

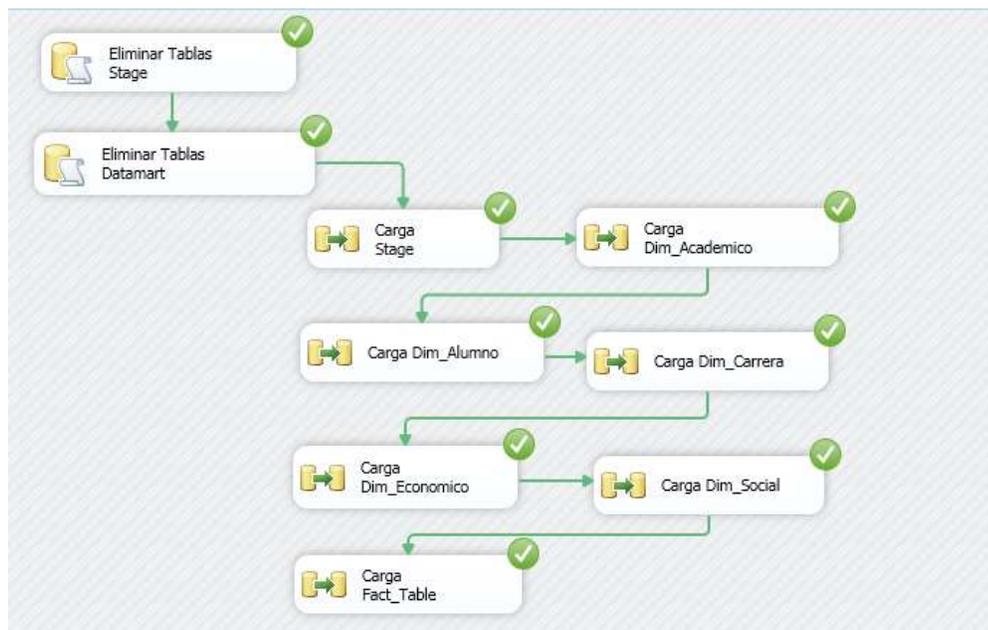


Ilustración 39: Carga ETL, Integracion Services

6.3 Modelos evaluados

El procesamiento de los datos se realizó en la herramienta de BI Analysis Services, el cual satisface los objetivos del modelo predictivo.

En la ilustración 40, podemos observar que se evaluaron 6 modelos de predicción, todos estos modelos fueron evaluados con las mismas variables de entrada y ser comparados mediante el grafico de elevación que nos muestra la herramienta.

Se observa el tipo de entrada de las variables escogidas, que variable se usa como llave y la variable que será precedida por el modelo.

Entre los modelos evaluados que se aprecian en la ilustración 40, tenemos los siguientes:

1. Árboles de Decisión
2. Clustering
3. Bayes Naive
4. Redes neuronales
5. Regresión logística
6. Reglas de Asociación

Estructura	Arboles	Clustering	Bayes naïve	Red Neuronal	Regresión	Reglas de asociación
	Microsoft_Decision_Trees	Microsoft_Clustering	Microsoft_Naive_Bayes	Microsoft_Neural_Network	Microsoft_Logistic_Regress...	Microsoft_Association_Rules
✓ CREDENCIENTES	Input	Input	Input	Input	Input	Input
✓ CURSOS/APROBADOS	Input	Input	Input	Input	Input	Input
✓ DISTRITO	Input	Input	Input	Input	Input	Input
✓ ESTADO	PredictOnly	PredictOnly	PredictOnly	PredictOnly	PredictOnly	PredictOnly
✓ ID/ALUMNO	Key	Key	Key	Key	Key	Key
✓ NIVEL/SOCIOECONOMICO	Input	Input	Input	Input	Input	Input
✓ SEXO	Input	Input	Input	Input	Input	Input

Ilustración 40: Modelos de Predicción Evaluados

En la ilustración 41, podemos observar que luego del procesamiento la herramienta Analysis Services califica la precisión de predicción de los modelos.

Se aprecia que la técnica de árboles de decisión tiene una mayor precisión con respecto a los otros modelos evaluados.

Serie, Modelo	Puntuación	Población correcta
Arboles	0.84	44.44%
Clustering	0.59	26.67%
Bayes naive	0.69	37.78%
Red Neuronal	0.72	40.00%
Regresión	0.70	40.00%
Reglas de asociaci...	0.60	28.89%
Modelo ideal		51.00%

Ilustración 41: Leyenda del Gráfico de Elevación

En la ilustración 42, se observa que los resultados sobre la precisión de cada modelo se plasman en un gráfico de elevación, el cual nos muestra como cada uno los modelos evaluados se van acercando al modelo ideal representado con una línea recta azul.

El eje X del gráfico representa el porcentaje del conjunto de datos de prueba que se usa para comparar las predicciones. El eje Y del gráfico representa el porcentaje de valores de predicción.

Según el color de la leyenda de la ilustración 41, se observa que el modelo arboles de decisión tiene mayor precisión en resultados debido que es el más cercado al modelo ideal, seguido por el modelo de Clustering.

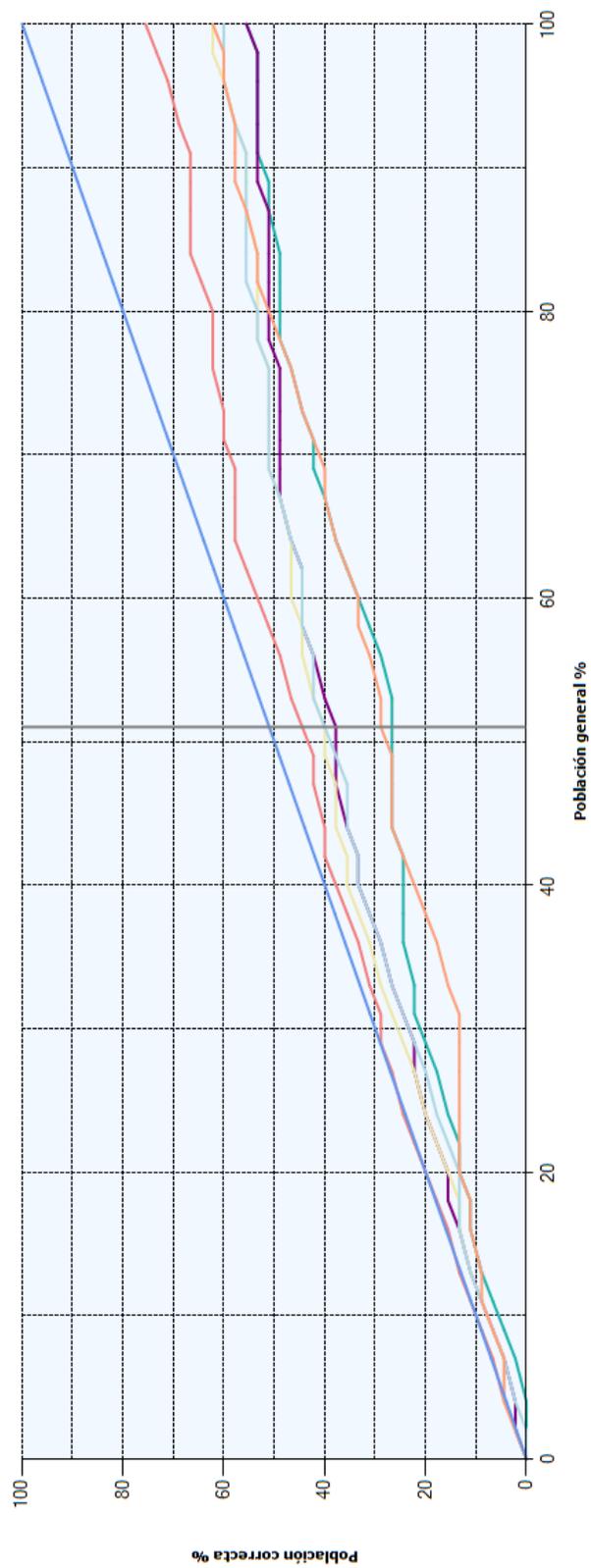


Ilustración 42: Gráfico de elevación de minería de datos

6.4 Resultados del Árbol de decisiones

Al realizar el análisis del resultado del modelo de árboles de decisión, se identifican los principales factores que desencadenan la deserción en los alumnos de la escuela informática.

En la ilustración 43, podemos observar que el color azul marca la concentración de deserción de los alumnos de la escuela de ingeniería informática.

Se aprecia que las variables que determinan la deserción de los alumnos están divididas en 3 niveles, teniendo en el primer nivel la variable más predominante en el resultado.

En el primer nivel del árbol, tenemos la cantidad de créditos pendientes y podemos observar la mayor concentración de deserción en los alumnos con créditos pendientes mayor a 200, es decir de los primeros ciclos.

En el segundo nivel del árbol, podemos observar como el nivel socioeconómico "C" concentra la mayor cantidad de desertores, convirtiéndose en un factor de deserción.

Como último y tercer nivel del árbol, tenemos la cantidad de cursos desaprobados, la mayor cantidad de desertores se concentra en los alumnos que desaprobaron más de 4 cursos.

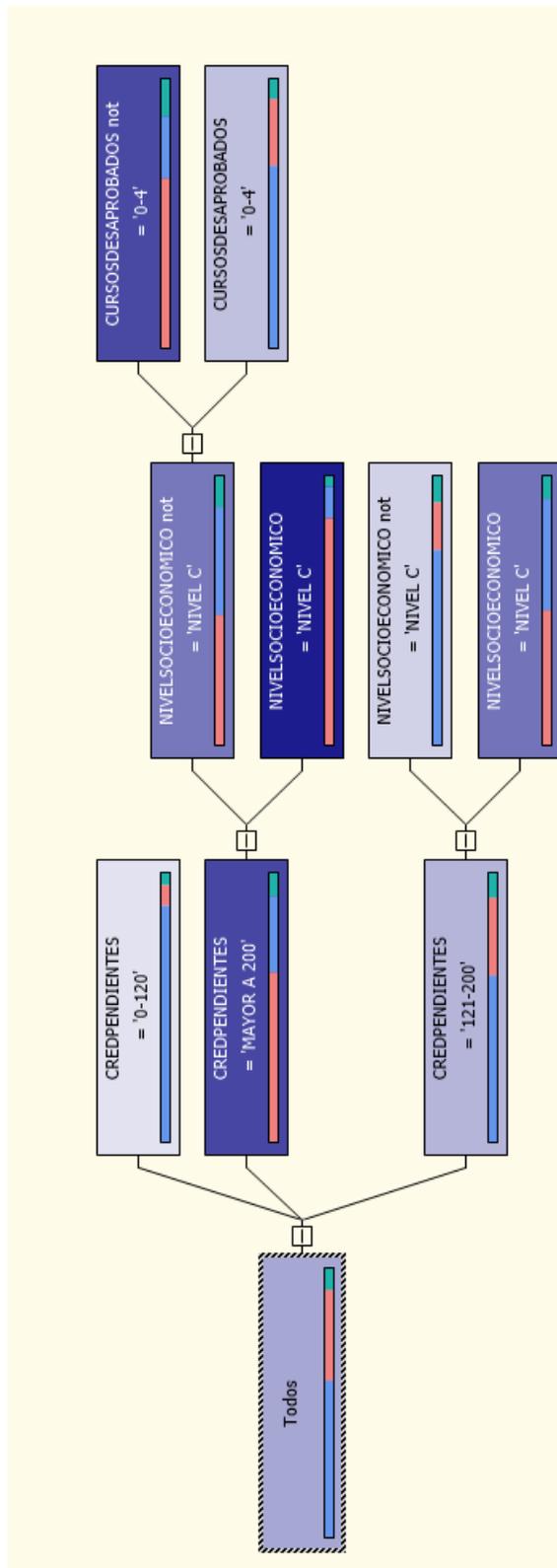


Ilustración 43: Árbol de decisión de la deserción universitaria en la escuela informática

Podemos concluir que los factores que determinan la deserción son los siguientes.

1. Económico

El nivel socioeconómico que concentra la cantidad mayor deserción de alumnos es nivel "C".

2. Académico

Influye la cantidad de créditos pendientes que determina que tan pronto está el alumno de culminar la carrera, también la cantidad de créditos desaprobados que refleja el rendimiento del alumno.

6.5 Perfil del desertor

Podemos deducir que el desertor de la carrera de ingeniería informática es un alumno perteneciente a los primeros ciclos de la carrera, se encuentra dentro del nivel socioeconómico C y durante su estadía en la universidad ha desaprobado más de 4 cursos.

Los factores antes mencionados nos permitirán identificar a alumnos activos que muestren las mismas características y pronosticar la probabilidad de deserción.

El perfil del desertor puede cambiar durante el tiempo, el modelo podrá madurar con la carga de nueva información y se podrán determinar nuevos factores que podrán ser aplicados a la predicción.

6.6 Producto

Se desarrolló una aplicación de fácil uso para que la Escuela informática pueda acceder a los resultados sobre la deserción en la carrera de Ingeniería informática

CONCLUSIONES

1. Las herramientas de BI utilizadas como el Analysis Services e Integration Services, satisficieron la realización de los requerimientos funcionales y no funcionales del proyecto.
2. Se revisaron diferentes técnicas de minería de datos para desarrollar modelos de predicción, tales como arboles de decisión y cluster.
3. Como resultado del modelo predictivo se puede observar que el factor económico y académico son los más predominantes en el problema de la deserción.
4. Según el estudio revisado en el “ANEXO 7”, refleja claramente que el nivel socioeconómico C tiene alta probabilidad de no poder pagar los gastos de pensión de la Universidad Ricardo Palma cuyo monto se encuentra sobre los 800 nuevos soles.
5. Al analizar las diferentes técnicas que se utilizan en la minería de datos específicamente para la predicción, se observa que los arboles de decisión resultan ser buenos clasificadores, según los resultados obtenidos.
6. Para la creación del modelo predictivo se tuvieron en cuenta variables de tipo personal como, distrito de procedencia, nivel socio económico, sexo; variables económicas como valor de la matricula e ingresos y de carácter académico como, cursos desaprobados, créditos pendientes.
7. Este trabajo permitió apreciar la importancia que tiene el proceso de recopilación de datos, abarcando las fases de análisis y preparación de los datos descrito en el capítulo 2 asociado a la metodología CRISP-DM.

8. El personal administrativo podrá usar la información que arroje el modelo para identificar y tomar acciones sobre los alumnos que serán posibles desertores y mejorar el índice de retención.

RECOMENDACIONES

1. Este trabajo permitió identificar que la universidad no recolecta la suficiente información referente a la caracterización del estudiante al momento de ingresar a estudiar, que permitan establecer modelos de predicción.
2. Para obtener un mejor resultado se deberá de trabajar con datos reales de los alumnos, para poder lograr los niveles de predicción necesarios para la validación positiva de los modelos.

BIBLIOGRAFIA

[1]Fundación Big Data (2015). *Análisis Predictivo*

Disponible en: <http://fundacionbigdata.org/analisis-predictivo/>

(Accesado 16 Julio 2015)

[2]Ing. Sonia Alejandra Formia, Universidad Nacional de La Plata (2012). *Evaluación de Técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios*

Disponible en:

http://sedici.unlp.edu.ar/bitstream/handle/10915/26772/Documento_completo.pdf?sequence=1

(Accesado 16 Julio 2015)

[3] Erika Himmel (2002). *Modelos de Análisis de la deserción estudiantil en la educación superior.*

Disponible en:

http://www.universidadtecnologica.net/tportal/portales/tp4964b0e1bk102/uploadlm/g/File/EducacionSuperior/desercion/1_%20DesercionE_Himmel.pdf

(Accesado 16 Julio 2015)

[4]Sandra Guzmán, Pontificia Universidad Javeriana (2009). *Deserción y retención estudiantil en los programas de pregrado de la Pontificia Universidad Javeriana*

Disponible en :<http://repository.javeriana.edu.co/bitstream/10554/425/1/edu54.pdf>

(Accesado 16 Julio 2015)

[5]Maria del Pilar Mori Sanchez, Universidad Peruana del Oriente (2012). *Deserción Universitaria en estudiantes de una Universidad Privada en Iquitos.*

Disponible en:<http://www3.upc.edu.pe/html/0/boletines/ridu/articulo-4-desercion-estudiantil-mori.pdf>

(Accesado 16 Julio 2015)

[6] Andrea Mirez Vargas, Universidad de Piura (2014). *La deserción estudiantil en la Facultad de Ciencias Económicas y Empresariales de la Universidad de Piura, y su impacto en los resultados económicos*

Disponible en:

http://pirhua.udep.edu.pe/bitstream/handle/123456789/1792/AE_268.pdf?sequence=1

(Accesado 16 Julio 2015)

[7] Vincent Tinto (1989). *Definir la deserción: Una Cuestión de Perspectiva*

Disponible en:

<http://preu.unillanos.edu.co/sites/default/files/fields/documentos/vicen%20tinto%20deser.pdf>

(Accesado 16 Julio 2015)

[8] José Hernández, María José Ramírez, and César Ferri, *Introducción a la Minería de Datos*. Madrid: Pearson Educación S. A., 2004.

[9] Pete Chapman et al. (1999) *CRISP-DM 1.0 Step-by-step data mining*.

Disponible en:

<ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf>

(Accesado 16 Julio 2015)

[10] Dr. Saed Sayad (2015). *An Introduction to Data mining*

Disponible en: http://www.saedsayad.com/data_mining_map.htm

(Accesado 16 Julio 2015)

[11] Asociación Peruana de Empresas de Investigación de Mercado (2014).

Estudio Niveles Socioeconómicos 2014

Disponible en <http://www.apeim.com.pe/wp-content/themes/apeim/docs/nse/APEIM-NSE-2014.pdf>

(Accesado 15 octubre 2015)

[12] Instituto Internacional para la Educación Superior en América Latina y Caribe (2009). *Estudio sobre la repitencia y deserción en la educación superior chilena*

Disponible en:

<http://www.iesalc.unesco.org.ve/>

(Accesado 15 octubre 2015)

[13] Ministerio de Educación Nacional de Colombia (2009). *Deserción estudiantil en la educación superior colombiana*

Disponible en:

http://www.mineduacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

(Accesado 15 octubre 2015)

[14] Logros (2012). *Deserción Universitaria*

Disponible en: <http://www.logrosperu.com/noticias/institucionales/775-desercion-universitaria.html>

(Accesado 15 octubre 2015)

[15] Flanqueo (2013). *Deserción de Alumnos*

Disponible en: <http://laprensa.peru.com/economia/noticia-abondono-estudiantes-universidades-disminuyen-hasta-30-sus-ingresos-6622>

(Accesado 15 octubre 2015)

[16] Luis Eduardo Gonzales (2005). *Repitencia y Deserción en América Latina*

Disponible en:

http://www.academia.edu/5891092/Repitencia_y_Deserci%C3%B3n_Universitaria_en_Am%C3%A9rica_Latina

(Accesado 22 agosto 2015)

[17] Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (2015). *América Latina y Caribe: Revisión Regional 2015 de la educación para Todos*

Disponible en:

<http://www.unesco.org/new/fileadmin/MULTIMEDIA/FIELD/Santiago/pdf/Informe-Regional-EFA2015.pdf>

(Accesado 22 agosto 2015)

ANEXO 1: Flujo alternativo de CUS Transformar y cargar datos

1. Error: Ejecución del Agente SQL

Si al hacer clic en el botón “Ejecutar cargar de información” y el Agente SQL no se encuentra iniciado, el sistema nos muestra el siguiente error como se muestra en la ilustración 44.

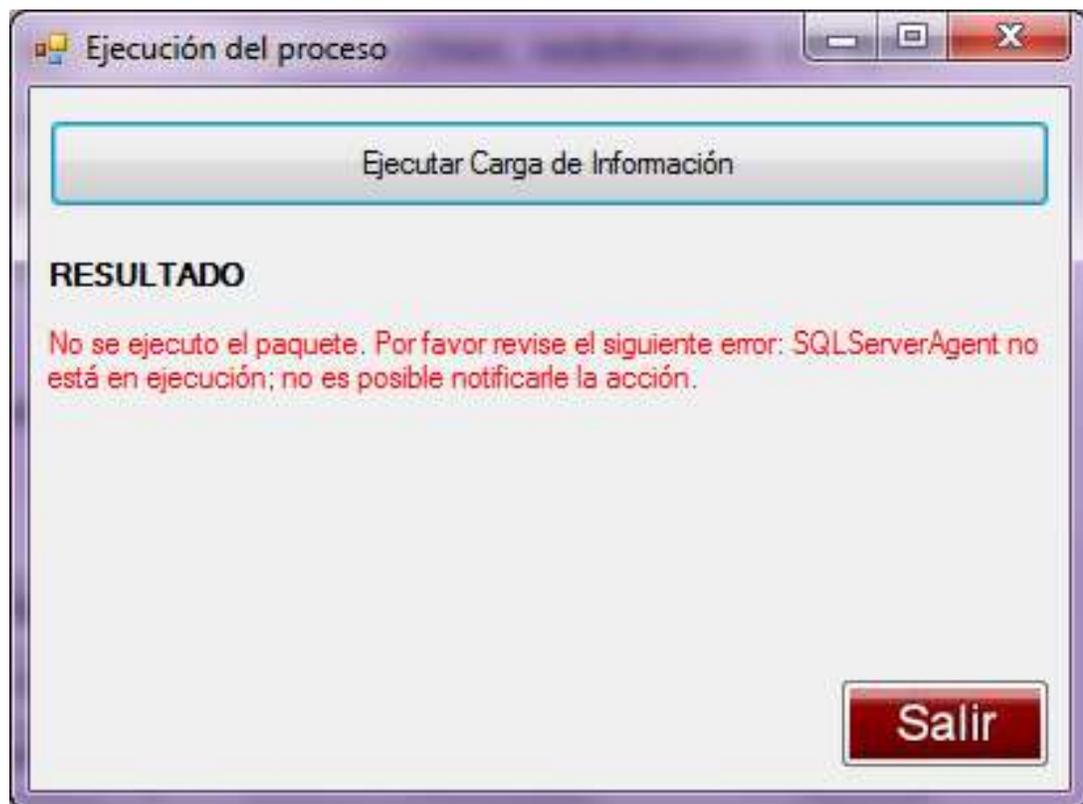


Ilustración 44: Error en la ejecución del Agente SQL

Ingresamos al SQL server y como se aprecia en la ilustración 45, observamos que el Agente SQL Server se encuentra detenido.

Clic derecho sobre el Agente SQL server y damos clic en la opción “iniciar”

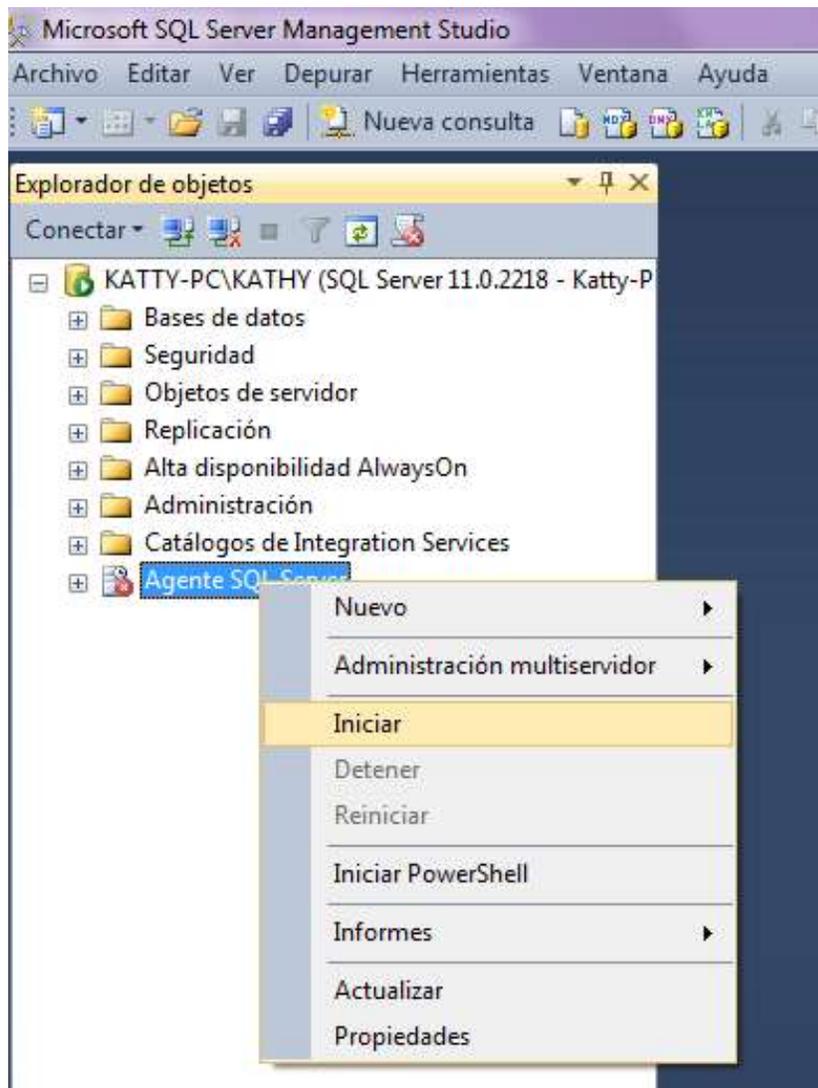


Ilustración 45: Agente SQL Server detenido

Como podemos apreciar en la ilustración 46, el Agente SQL Server se encuentra iniciado.

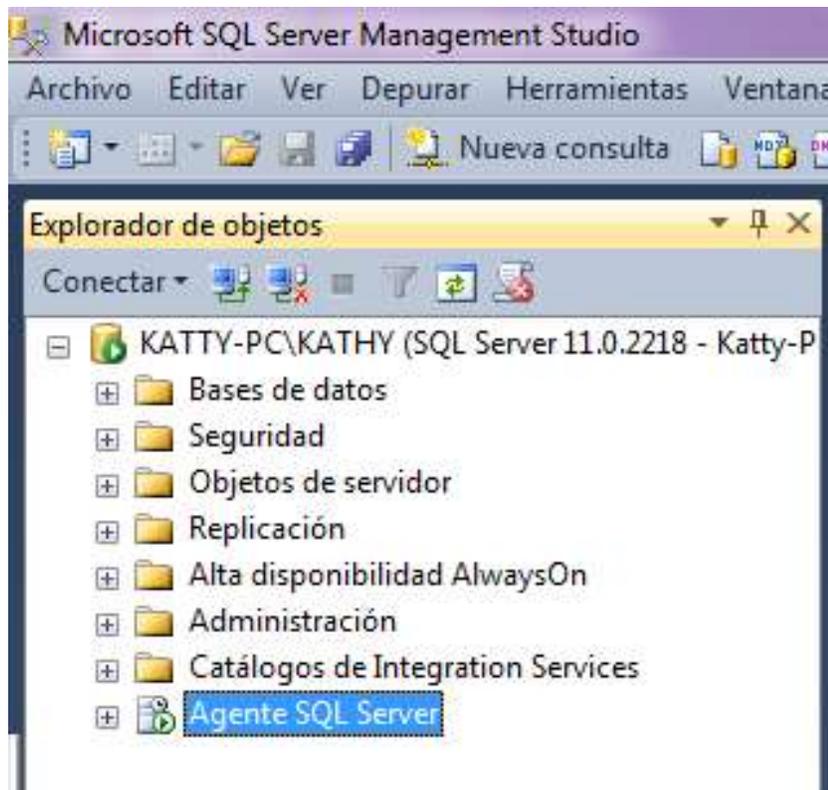


Ilustración 46: Agente SQL Server Iniciado

2. Error: Conectar al Servidor SQL Server

Si al hacer clic en el botón “Consultar Datos” y el servicio del servidor SQL server no se encuentra iniciado, el sistema nos muestra el siguiente error como se muestra en la ilustración 47.

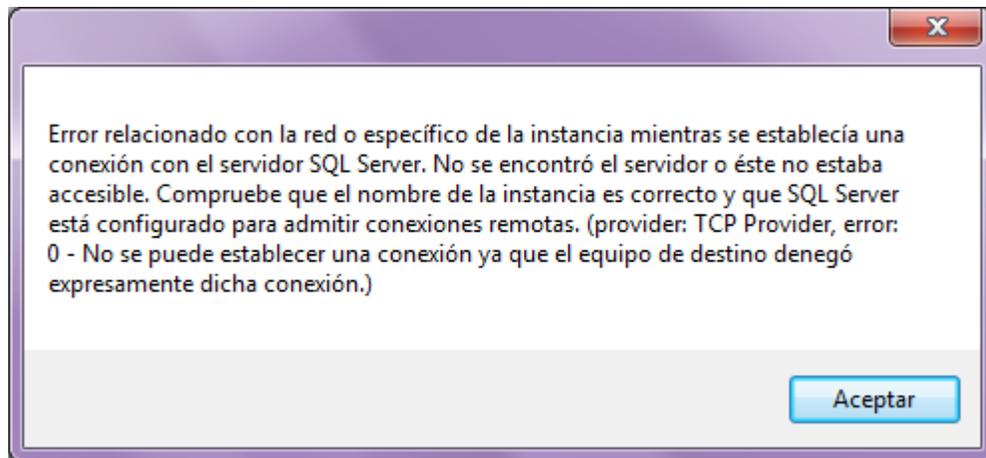


Ilustración 47: Error en el inicio de Servicio SQL

Ingresamos a los servicios del sistema y como se aprecia en la ilustración 48, observamos que el servicio SQL Server se encuentra detenido.

Seleccionamos el servicio SQL server y damos clic en la opción “iniciar”

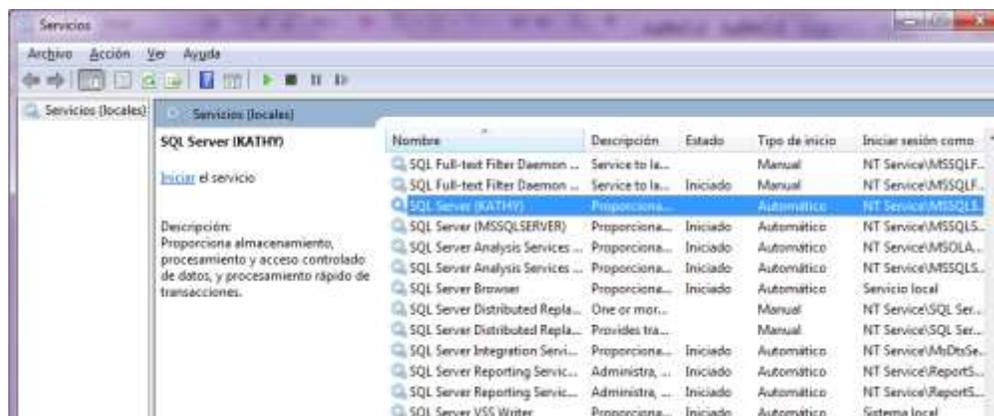


Ilustración 48: Servicio SQL Server Iniciado

ANEXO 2: Informe de Variables

1. Análisis de datos

La fase de análisis de datos ha comprendido la recolección de los datos de los estudiantes de la facultad de Ingeniería Informática de la Universidad Ricardo Palma, obtenidos desde OFICIC. La selección de las variables a utilizar para este estudio se basa principalmente atendiendo los factores descritos en Tabla 4.

Los atributos con sus descripciones y posibles valores son presentados en la Tabla 15. El atributo “Estado” es el indicador de deserción de los estudiantes de Ingeniería Informática”.

Tabla 15: Atributos usados en el modelo predictivo

Atributo	Descripción	Posibles Valores
Aleatorio	Identificador del alumno	Número
Sexo	Género	M(Masculino), F(Femenino)
Distrito	Distrito de Residencia del estudiante	Todos los distritos
Nivel Socio Económico	Nivel socio económico del estudiante	Nivel A, Nivel B, Nivel C
Cursos desaprobados	Historial de cursos desaprobados del alumno	Mayor o igual a 0
Créditos pendientes	Créditos pendientes del alumno	Rango 0 a 220
Estado	El estado a predecir	Activo, Desertor

ANEXO 3: Cronograma del Proyecto

A continuación en la tabla 16, se muestran tareas y actividades realizadas durante el ciclo de vida del proyecto de tesis.

Tabla 16: Cronograma del Proyecto

	Trabajo	Duración
MODELO DE DESERCIÓN DE ESTUDIANTES	420 HORAS	140 DÍAS
Identificación con la empresa	45 horas	15 días
Definición del tema		
Identificación de Problemática y Objetivos		
Metodología a emplear		
Comprensión del negocio	45 horas	15 días
Objetivos del Negocio		
Requerimientos		
Restricciones		
Riesgos y Costos		
Beneficios		
Objetivos de la minería de datos		
Plan de proyecto		
Comprensión de los datos	45 horas	15 días
Informe de colección de datos iniciales		
Descripción de los datos		
Informe de calidad de datos		
Preparación de los datos	60 horas	20 días
Informe de selección de datos		
Limpieza de datos		
Modelado	150 horas	50 días
Selección de técnica de modelado		
Pruebas de diseño		
Evaluación de modelos		

Evaluación	30 horas	10 días
Revisión del proceso		
Despliegue	45 horas	15 días
Plan de Desarrollo		
Informe final		

ANEXO 4: Estrategia de Ejecución

En la ilustración 49, se observa la estrategia de ejecución que se llevó a cabo en esta tesis.

Nro.	CUS	PREDECESOR	Fecha de inicio	Fecha de Fin	DIAS	Complejidad
1	Proceso ETL Limpieza de Datos información	-	22-ago	27-ago	5	20
2	Actualizar documento de tesis	1	22-ago	23-ago		
3	Verificar Información Construcción del Modelo	2	24-ago	26-ago		
4	Actualizar documento de tesis	3	27-ago	27-ago		
5	Evaluar el modelo Actualizar documento de tesis	4	28-ago	30-ago	2	8
	Crear Informes de predicción Actualizar documento de tesis		05-sep	18-sep	13	78
			18-sep	18-sep		
			19-sep	24-sep	5	25
			24-sep	24-sep		
			25-sep	01-oct	6	20
			01-oct	01-oct		

Ilustración 49: Estrategia de Ejecución del modelo predictivo

ANEXO 5: Informe de Colección de Datos

En la tabla 17, se observa el informe de colección de datos obtenidos de OFICIC.

Fuente: OFICIC

Cantidad de Registros: 4,763

Fecha de Registros: 2011-1 al 2015-1

Tabla 17: Informe de Colección de Datos

ALEATORIO	CRED_ELEC_APROB
PERIODO	SITUACION
CARR-SEMESTRE	SEXO
CURSOSAPROBADOS	FECHA NACIMIENTO
CURSOSDESAPROBADOS	LUGAR NACIMIENTO
CURSOSNSP	DISTRITO NACIMIENTO
CREDAPROBADOS	ESTADO_CIVIL
CREDESAPROBADOS	DISTRITO DOMICILIO
CREDNSP	PROV COLEGIO
PROMEDIOSINNSP SEMESTRE	DISTR COLEGIO
SITUACION_ACTUAL_CARRERA	COLEGIO
SEM INGRESO	GRADO_INSTRUCCION_PADRE
SEM_NOTAS_HIST_INICIAL	PROFESION_PADRE,
PERIODOS_NOTAS_HIST	INGRESO_NETO_PADRE
PERIODOS_NOTAS_HIST_SIN_VERANO	INGRESOS_OTROS_PADRE
CRED_CURS_HIST_SIN_NSP	INGRESO_NETOS_OTROS_PADRE
SEM_MAT_HIST_INICIAL	GRADO_INSTRUCCION_MADRE
PERIODOS_MAT_HIST	PROFESION_MADRE
PERIODOS_MAT_HIST_SIN_VERANO	INGRESO_NETO_MADRE
CARR_ACTUAL	INGRESOS_OTROS_MADRE
PLAN_ACTUAL	INGRESO_NETOS_OTROS_MADRE
PROM_HIST	ESCA_NOMBRE
CREDAPROBADOS	ESCA_MONTO
CREDITOSCONVALIDADOS	BECA_CODIGO
CREDTOTALAPROBADOS	BECA_MOTIVO
CRED_CONV_OBLIG_APROB	PORCENTAJE
CRED_CONV_ELEC_APROB	NMRO_APLCDAS
CRED_OBLIG_APROB	RESTIRO_SEM

ANEXO 6: Informe de Calidad de Datos

En la tabla 18, podemos observar el informe inicial sobre la cantidad de registros por variable de la data total entregadas por OFICIC y que fue usada en el modelo predictivo de deserción. En la revisión de datos se encontraron las siguientes falencias

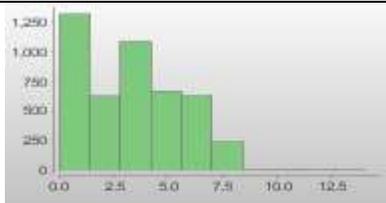
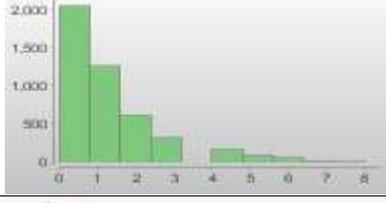
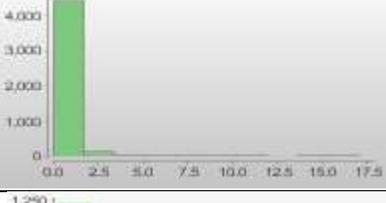
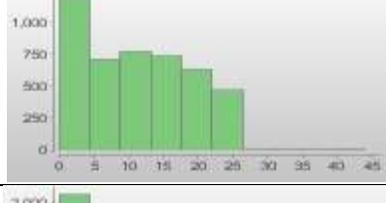
Tabla 18: Informe de registros por variable

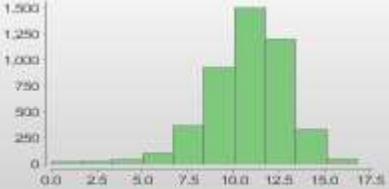
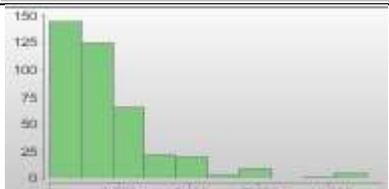
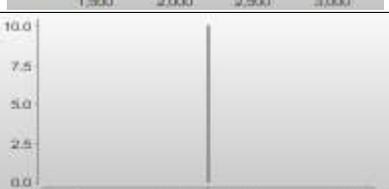
Variable	Total
LUGAR NACIMIENTO	58%
DISTRITO NACIMIENTO	90%
ESTADO_CIVIL	32%
DISTRITO DOMICILIO	96%
PROV COLEGIO	88%
DISTR COLEGIO	88%
COLEGIO	90%
GRADO_INSTRUCCION_PADRE	8%
PROFESION_PADRE,	99%
INGRESO_NETO_PADRE	8%
INGRESOS_OTROS_PADRE	56%
INGRESO_NETOS_OTROS_PADRE	8%
BECA_CODIGO	2%
BECA_MOTIVO	2%

1. Las variables de distrito presentaban caracteres que no pertenecen al alfabeto
2. No había un estándar en el nombre de los distritos
3. No se pudieron obtener al 100% datos sociales, económicos y demográficos. Las siguientes variables no presentaban campos vacíos o nulos.

Se realizó un análisis a las variables numéricas, como se puede observar en la tabla 19, el cual nos permitió detectar en que variables tenían valores extremos

Tabla 19: Análisis de variables numéricas

Variable	Valor min.	Valor máx.	Valor Prom.	Histograma
CURSOS APROBADOS	0	14	3.24	 <p>Este histograma muestra la distribución de los cursos aprobados. El eje horizontal representa el número de cursos (de 0.0 a 12.5) y el eje vertical representa la frecuencia (de 0 a 1,250). La distribución es unimodal y ligeramente sesgada a la izquierda, con un pico principal entre 0 y 2.5.</p>
CURSOS DESAPROBADOS	0	8	1.0	 <p>Este histograma muestra la distribución de los cursos desaprobados. El eje horizontal representa el número de cursos (de 0 a 8) y el eje vertical representa la frecuencia (de 0 a 2,000). La distribución es unimodal y altamente sesgada a la izquierda, con un pico muy alto en el primer bin (0-1).</p>
CURSOS NSP	0	17	0.18	 <p>Este histograma muestra la distribución de los cursos NSP. El eje horizontal representa el número de cursos (de 0.0 a 17.5) y el eje vertical representa la frecuencia (de 0 a 4,000). La distribución es unimodal y altamente sesgada a la izquierda, con un pico muy alto en el primer bin (0-2.5).</p>
CRED APROBADOS	0	44	11.0	 <p>Este histograma muestra la distribución de los créditos aprobados. El eje horizontal representa el número de créditos (de 0 a 45) y el eje vertical representa la frecuencia (de 0 a 1,250). La distribución es unimodal y sesgada a la izquierda, con un pico principal entre 0 y 5.</p>
CRED DESAPROBADOS	0	22	4.0	 <p>Este histograma muestra la distribución de los créditos desaprobados. El eje horizontal representa el número de créditos (de 0 a 20) y el eje vertical representa la frecuencia (de 0 a 2,000). La distribución es unimodal y altamente sesgada a la izquierda, con un pico muy alto en el primer bin (0-2).</p>
PROMEDIO SIN NSP	0	99	11.82	 <p>Este histograma muestra la distribución del promedio sin NSP. El eje horizontal representa el promedio (de 0 a 100) y el eje vertical representa la frecuencia (de 0 a 3,000). La distribución es unimodal y altamente sesgada a la izquierda, con un pico muy alto en el primer bin (0-25).</p>

PERIODO MAT HIST	1	33	9.0	
PROM HIST	0	17	10.7	
CRED CONVALIDADOS	0	185	5.0	
INGRESO NETO PADRE	0	11285	1984.8	
INGRESO OTROS PADRE	0	123123	455.7	
INGRESO NETO MADRE	1200	3200	1534.6	
INGRESO OTROS MADRE	0	0	0	
ESCA MONTO	795	1631	1007.3	

ANEXO 7: Variable Socio Económica

La variable nivel socioeconómico fue calculada a partir del distrito de domicilio, la escala del alumno y el ingreso neto de sus padres.

Según el estudio del 2014 de perfil de hogares según nivel socioeconómico de Lima Metropolitana de la Asociación Peruana de Empresas de Investigación de Mercados , el nivel socioeconómico C tiene un ingreso promedio de S/. 3,422.

El estudio señala también que se invierte en promedio un 12.6% en Educación, es decir un aproximado de 431 nuevos soles destinados a la educación.

En la ilustración 50, podemos observar cuanto son los gastos promedios de cada nivel socio económico.

PROMEDIOS	Lima Metropolitana							
	TOTAL	NSE A	NSE B	NSE C	NSE C1	NSE C2	NSE D	NSE E
Grupo 1 : Alimentos – gasto promedio	S/. 701	S/. 982	S/. 866	S/. 720	S/. 735	S/. 696	S/. 572	S/. 503
Grupo 2 : Vestido y Calzado – gasto promedio	S/. 96	S/. 275	S/. 135	S/. 89	S/. 93	S/. 81	S/. 62	S/. 53
Grupo 3 : Alquiler de vivienda, Combustible, Electricidad y Conservación de la Vivienda – gasto promedio	S/. 242	S/. 608	S/. 392	S/. 220	S/. 234	S/. 197	S/. 153	S/. 99
Grupo 4 : Muebles, Enseres y Mantenimiento de la vivienda – gasto promedio	S/. 96	S/. 663	S/. 134	S/. 59	S/. 63	S/. 52	S/. 43	S/. 39
Grupo 5 : Cuidado, Conservación de la Salud y Servicios Médicos – gasto promedio	S/. 144	S/. 364	S/. 219	S/. 137	S/. 155	S/. 109	S/. 92	S/. 58
Grupo 6 : Transportes y Comunicaciones – gasto promedio	S/. 218	S/. 1,170	S/. 392	S/. 154	S/. 179	S/. 114	S/. 75	S/. 50
Grupo 7 : Esparcimiento, Diversión, Servicios Culturales y de Enseñanza – gasto promedio	S/. 262	S/. 1,026	S/. 484	S/. 219	S/. 254	S/. 162	S/. 98	S/. 66
Grupo 8 : Otros bienes y servicios – gasto promedio	S/. 102	S/. 295	S/. 140	S/. 95	S/. 101	S/. 85	S/. 64	S/. 59
PROMEDIO GENERAL DE GASTO FAMILIAR MENSUAL	S/. 2,985	S/. 7,901	S/. 4,335	S/. 2,785	S/. 2,962	S/. 2,513	S/. 1,908	S/. 1,424
PROMEDIO GENERAL DE INGRESO FAMILIAR MENSUAL*	S/. 3,718	S/. 11,395	S/. 5,519	S/. 3,422	S/. 3,647	S/. 3,077	S/. 2,167	S/. 1,466

* Ingreso estimado en base al gasto - INEI

APEIM 2014: Data ENAHO 2013

Ilustración 50: Ingreso estimado por NSE en Lima Metropolitana

En la ilustración 51, podemos observar la distribución de gastos de diferentes rubros para cada nivel socio económico.



DISTRIBUCIÓN DEL GASTO SEGÚN NSE 2014 - LIMA METROPOLITANA

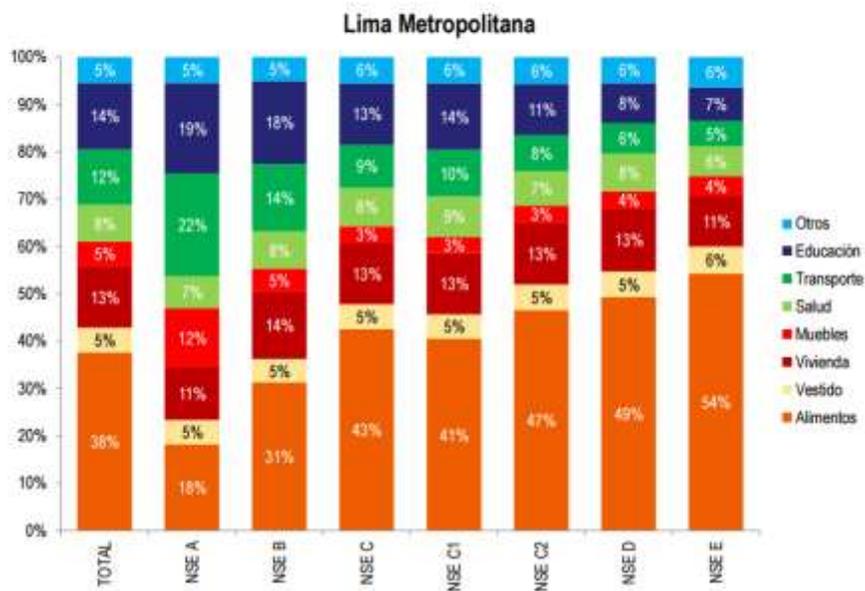


Ilustración 51: Distribución del Gasto por NSE en Lima Metropolitana

ANEXO 8: Instrucciones de Limpieza de Datos

Se tomaron las siguientes acciones para la limpieza de los datos con la herramienta SQL Server Integration Services:

1. Se calculó por alumno el total de créditos aprobados, créditos desaprobados, cursos Aprobados, cursos Desaprobados y créditos Pendientes.
2. Se sumaron los ingresos netos y otros de los padres en una sola variable.
3. Se calculó el nivel socioeconómico a partir del total ingreso padres, escala, beca, distrito de domicilio (previamente estandarizado)
4. Se calculó el estado de alumno considerando desertor al estudiante aun con créditos pendientes y que no se haya matriculado en los últimos 3 ciclos.
5. Se le dio un valor promedio a todos los valores nulos.
6. Se detectaron los valores extremos y fueron retirados.

ANEXO 9: Variables usadas en Modelo Predictivo

1. Modelo de Predicción

Se trabajó en la herramienta de BI Analysis Services en la cual se conecta a la Base de datos del proyecto BD_DESERCION y se tomaron los siguientes atributos de ingreso que se muestran en la ilustración 52.

Atributos: Créditos Pendientes, Cursos Desaprobados, Distrito, Estado, Código de Alumno, Nivel Socioeconómico y Sexo.

La variable que modelo va a predecir es “Estado”.

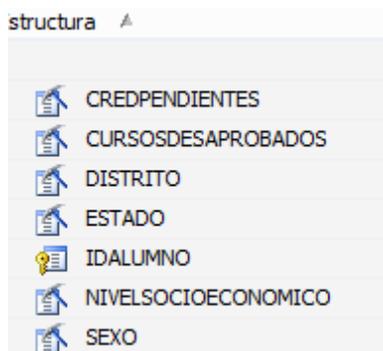


Ilustración 52: Variables del Modelo de Predicción

ANEXO 10: Confirmación de Carga de Datos

Al finalizar la ejecución del CUS: Transformar y Cargar datos, el sistema envía un correo electrónico informando el estado correcto de la carga de información de los alumnos a la base de datos del proyecto.

En la ilustración 53, podemos observar la notificación de correo enviada por el sistema.



Ilustración 53: Correo electrónico de confirmación de ejecución del CUS: Transformar y Cargar datos