

UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR
FACULTAD DE INGENIERÍA Y GESTIÓN
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS



**“DISEÑO DE UN MODELO PREDICTIVO BASADO EN MACHINE
LEARNING PARA EL CONTROL DE LA DESERCIÓN DE
ESTUDIANTES EN LA UNIVERSIDAD RICARDO PALMA”**

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de

INGENIERO DE SISTEMAS

**PRESENTADO POR EL BACHILLER
DE LA CRUZ QUISPE, VÍCTOR RAUL**

Villa el Salvador

2019

DEDICATORIA

A Dios, por darme la oportunidad de llegar hasta este punto de mi vida universitaria y haberme dado salud cuando más lo necesitaba para lograr mis objetivos.

A mi familia, por darme una carrera para mi futuro, por aconsejarme y estar conmigo en los momentos más difíciles.

Finalmente, a mis amigos, porque de cada uno aprendí muchas cosas valiosas, por haber compartido buenos y malos momentos a lo largo de nuestra vida universitaria.

AGRADECIMIENTO

Quiero agradecer a mi alma mater, por acogerme en estos años de formación profesional.

También agradecer a cada uno de los docentes, por su dedicación y profesionalidad. Ha sido fundamental para mi desarrollo como profesional.

ÍNDICE

DEDICATORIA	ii
AGRADECIMIENTOS	iii
INTRODUCCIÓN	1
CAPÍTULO I:	3
PLANTEAMIENTO DEL PROBLEMA	3
1.1. Descripción de la Realidad Problemática	3
1.2. Justificación	5
1.3. Delimitación del Proyecto	7
1.3.1. Temporal	7
1.3.2. Espacial	7
1.3.3. Conceptual	7
1.4. Formulación del Problema	8
1.4.1. Problema General	8
1.4.2. Problemas Específicos	8
1.5. Objetivos	9
1.5.1. Objetivo General	9
1.5.2. Objetivos Específicos	9
CAPÍTULO II:	10
MARCO TEÓRICO	10
2.1. Antecedentes	10
2.2. Bases Teóricas	18
2.2.1 Metodología para minería de datos (CRISP-DM)	18
2.2.2 Data Mining	21
2.2.3 Análisis Predictivo	21
2.2.1 Machine Learning	22
2.2.2 Técnicas de Machine Learning	23
2.2.3 Deserción Universitaria	24

2.2.4 Big Data.....	28
2.2.5 Hadoop.....	29
2.2.6 Metodología Scrum.....	29
CAPÍTULO III:.....	32
DESARROLLO DEL TRABAJO DE SUFICIENCIA PROFESIONAL	32
3.1 Modelo de Solución Propuesto	32
- Entendimiento del negocio.....	32
- Preparación de datos	38
- Modelado.....	40
- Evaluación.....	56
- Despliegue del modelo	57
3.2 Resultados.....	77
CONCLUSIONES.....	82
RECOMENDACIONES	84
BIBLIOGRAFÍA.....	85

LISTADO DE FIGURAS

Figura 1: Cantidad de estudiantes matriculados	4
Figura 2: Porcentaje de estudiantes de deserción URP	5
Figura 3: Niveles de la metodología CRISP-DM	19
Figura 4: Instalación Sql Server 2016	33
Figura 5: Instalación de Visual Studio.....	34
Figura 6: Restauración BD URP	35
Figura 7: Esquema del Datamart	36
Figura 8: Creación del Datamart.....	37
Figura 9: Ejecución del Datamart.....	38
Figura 10: Nueva carga de Datamart.....	39
Figura 11: Carga BD transaccional	39
Figura 12: Interfaz Big Data HDFS.....	40
Figura 13: Características de Servidor HDFS.....	41
Figura 14: Componentes de HDFS	42
Figura 15: File System HDFS.....	42
Figura 16: Data recompilada y Procesada del Datamart	43
Figura 17: Excel Exportado a HDFS	44
Figura 18: Aplicación Open Source Instalada R y Rstudio.....	44
Figura 19: Librerías del modelo predictivo	45
Figura 20: Extracción de Data.....	46
Figura 21: Exploración de la Data	47
Figura 22: Imputación de la Data.....	48
Figura 23: Particionamiento de la data	49
Figura 24: Algoritmo Boruta.....	50
Figura 25: Algoritmo Naive Bayes	51
Figura 26: Algoritmo de Regresión Logístico.....	52
Figura 27: Algoritmo Árbol Chaid.....	53
Figura 28: Algoritmo Árbol Cart.....	53
Figura 29: Algoritmo Árbol c5.0.....	54
Figura 30: Algoritmo SVM Radial.....	54
Figura 31: Algoritmo SVM Linear.....	55
Figura 32: Resultado de los Modelos.....	56

Figura 33: Modelo Predictivo Final	57
Figura 34: SCRUM y Jira	61
Figura 35: Requerimientos Realizados.....	62
Figura 36: Diseño de los requerimientos.....	63
Figura 37: Prototipo Login.....	67
Figura 38: Prototipo de reporte estudiantes.....	68
Figura 39: Prototipo carga de Datamart	69
Figura 40: Prototipo Modelo Predictivo.....	70
Figura 41: Prototipo Dashboard	71
Figura 42: Modelo Físico de Base de Datos Utilizada (Datamart)	72
Figura 43: Módulo Login Aplicación	73
Figura 44: Módulo Reporte de estudiantes	74
Figura 45: Módulo Carga de Datamart	75
Figura 46: Módulo Carga Modelo Predictivo.....	76
Figura 47: Listado de Desertores	77
Figura 48: Principales factores de deserción.....	78
Figura 49: Deserción General de Estudiantes URP	79

LISTADO DE TABLAS

Tabla 1: Factores que impactan en la Deserción	25
Tabla 2: Tabla de Prioridades.....	58
Tabla 3: Requerimientos del sistema (Product backlog).....	58
Tabla 4: Designación de Tareas	64
Tabla 5: Porcentajes de deserción por total de desertores.....	80
Tabla 6: Porcentaje deserción por total de Alumnos.....	81

INTRODUCCIÓN

Este trabajo de investigación presenta un modelo predictivo inteligente que aprende automáticamente de acuerdo a la información que se le brinda, estos tipos de algoritmos de Machine Learning utilizan los datos de tal manera que puedan predecir comportamientos futuros automáticamente. Toda esta metodología implica que los sistemas creados mejoren de manera autónoma con el pasar del tiempo. Nuestro proyecto identificará el perfil de deserción estudiantil en la Universidad Ricardo Palma. Esto beneficia directamente al estudiante y a la institución, el objetivo es que los estudiantes deben concluir con su formación académica.

Actualmente, la Universidad tiene una alta tasa de deserción que puede alcanzar el 22% en los primeros ciclos (Ciclo 2-6), cuando se realiza la transición de los cursos básicos a los de especialización. Esto ocasiona muchos problemas en la programación académica de la respectiva Facultad.

Se ha hecho un análisis que estos indicadores son a raíz del bajo nivel académico, no adecuación del perfil del estudiante, desinterés por parte de los profesores, mal atención el proceso de matrícula y asignación de horarios de los cursos, alto costo de la matrícula, y un supuesto problema de una falta de selección de los potenciales estudiantes e inadecuada atención a los estudiantes.

En el Capítulo I, se describe detalladamente el trabajo realizado, es decir cómo se han implementado los objetivos generales y específicos.

Primero se describirá la problemática general por lo cual se realiza el proyecto, posteriormente se detallará la justificación, la delimitación del proyecto, el problema de la investigación y los objetivos.

En el Capítulo II, se describe el concepto de Machine Learning, Big Data, Hadoop, HDFS, deserción, entre otros. También se detallará los trabajos anteriores a este, los cuales se tomaron de referencia y los términos básicos correspondientes a los términos tecnológicos utilizados.

En el Capítulo III, se presentará los modelos de la solución y resultados obtenidos sobre la implementación de una solución de análisis predictivo. Por lo cual se tiene que tener acceso a la data de la Universidad para poder realizar un Datamart para filtrar la data, posteriormente toda esta Base de Datos deberá ser almacenada en una Base de Datos no relacional Hive, que será montada en un servidor Big Data con la tecnología HDFS. Luego se tendrá que realizar el modelo

predictivo basado en Machine Learning que nos predecirá los estudiantes que tienden a retirarse de la universidad y por lo cual con estos datos se deberá mostrar la información en unos sistemas de reportes con la predicción.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción de la Realidad Problemática

El problema por la cual se está realizando esta investigación, es la alta tasa de deserción estudiantil general que ocasiona pérdidas a las instituciones al no recabar por concepto de pago de estudios de los estudiantes retirados, y a la problemática general de como vallan a terminar dichos estudiantes, si cambian de universidad, si dejan de estudiar entre muchos más.

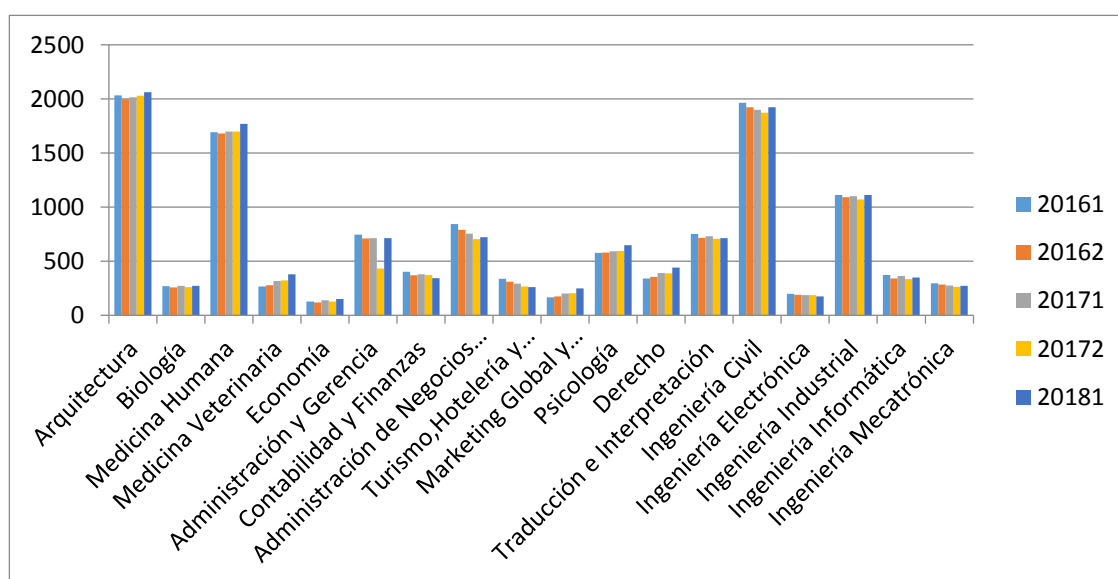
Dentro del análisis tomaremos a la Universidad Ricardo Palma para fuente de nuestros datos y solucionando su problema ya que estudios generales proporcionados por el área de sistemas OFICIC indica que tomando en cuenta desde los ciclos 2016-I hasta el 2018-II, y haciendo un análisis por carreras, la carrera de Turismo, Hotelería y Gastronomía tiene una deserción de 22% que son 75 estudiantes, la segunda carrera con mayor tasa de deserción es Contabilidad y Gerencia con un 15% que representa 59 estudiantes y seguido de la carrera de Administración de Negocios Globales, con un 14% que representa a 121 estudiantes todas del total de los estudiantes del ciclo 2016-I en comparación. Toda esta información ha sido recopilada de la Base de Datos de la Universidad por lo cual los altos directivos se han mostrado muy preocupados ya que se muestran facultades dentro de la Universidad vacías, profesorado contratado con pocos estudiantes y los estados financieros en rojos de la Universidad. Esta problemática se ha venido dando desde los últimos años, realizando un análisis esto comienza a darse debido a las nuevas universidades particulares que han venido creándose en el Perú, especialmente en Lima por lo cual la competitividad de las universidad por retener o aumentar las cantidades de los s nuevos y regulares es mayor.

Cabe recalcar también que algunos de los ámbitos de la salida de estudiantes son debido a la amplia gama de centros de estudio que le da al estudiante la facilidad de escoger de acuerdo a estado financiero y contextual. Analizando internamente puede llevarnos a múltiples

problemáticas como, las notas de los alumnos, sus asistencias, su afinidad con las carreras, el pago que deben realizar entre otras.

En la figura 1, se observa un gráfico de barras con la cantidad de alumnado que había ciclo tras ciclo. Información brindada por las Oficina Central de Computo (OFICIC).

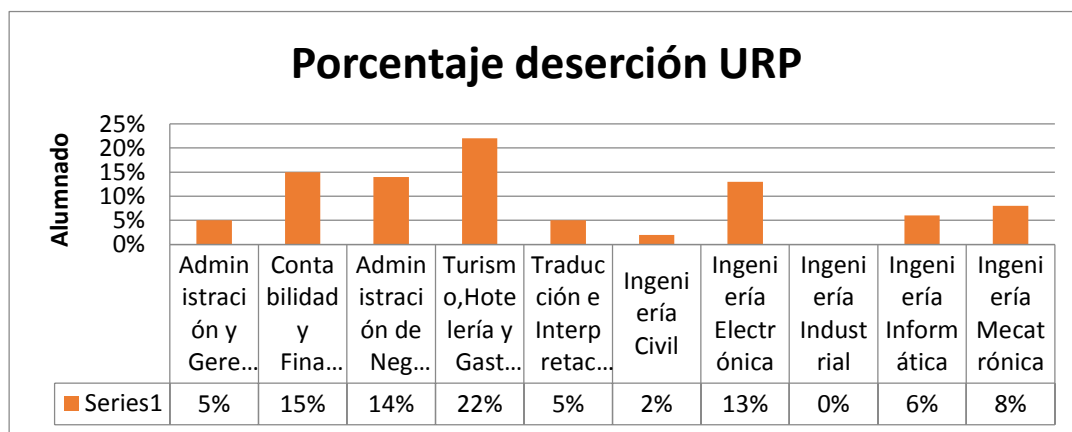
Figura 1: Cantidad de estudiantes matriculados



Fuente: Elaboración Propia

En la figura 2, se observa porcentajes de la deserción en comparación con el ciclo 2015-I especialmente de las carreras con más alto porcentaje de la deserción.

Figura 2: Porcentaje de estudiantes de deserción URP



Fuente: Elaboración Propia

1.2. Justificación

Este trabajo de investigación permitirá a la Universidad Ricardo Palma identificar a los estudiantes que serán desertores, de tal forma que pueda tomar acciones que prevengan el abandono de la carrera o cambio hacia otra universidad.

Ante la problemática anteriormente descrita, se realizará la metodología CRISP-DM la cual estructura los procesos de minería de datos en seis fases, las cuales interactúan entre ellas mismas, por lo cual se realizarán varios modelos predictivos en Machine Learning para que puedan ser analizados y luego de buscar el mejor identificarlo algoritmo que mejor convenga utilizarlo. Los algoritmos a utilizar serán las más seguras y con unas altas tasas de aceptación en la mayoría de empresas que realizan trabajos de predicción. Cabe indicar que el motivo de comprar muchos algoritmos es que debe ser modular y adaptarse a diferente tipo de información, es decir que al probarlo en otra data, quizás de otras universidades puede funcionar bien. Es por eso que el algoritmo ganador,

será utilizado en con la información de Universidad Ricardo Palma por lo cual se tomará mucho en cuenta la sensibilidad de los modelos. También nos determinara las características principales de los estudiantes desertores, todo este algoritmo que planteamos trabaja en un ambiente emulado de Big Data en un clúster de Hadoop que se denomina HDFS, con esta solución planteada se utilizará el software estadístico R y así podemos tener un mejor aprovechamiento de la información que será obtenido por un Datamart que es obtenido por la BD actual de la URP. Se implementó un clúster en Hadoop (HDFS) para que se pueda trabajar con grandes cantidades de información de manera eficiente y rápida, debido a que posee una tecnología de distribución simultánea.

Con esta solución en la Universidad Ricardo Palma ya no necesitará usar herramientas como excel, ni las estadísticas tradicionales ya que nuestro algoritmo predice los estudiantes a retirarse. Esto ayudará al análisis que realiza el área de Oficina Central de Cómputo de la Universidad Ricardo Palma correspondiente a la deserción.

La aplicación nos brindará varios módulos que representaran el proyecto completo automatizado, desde la recolección de la información con el Datamart hasta los reportes de los estudiantes que según nuestro modelo se retiraran de la Universidad Ricardo Palma.

Esta información es entregada a la Universidad Ricardo Palma en Bienestar Universitario que es uno de sus áreas internas que verifican este tipo de casos. Luego ellos deberían tomar las estrategias necesarias para disminuir dichas deserciones, mejorando los ámbitos que son como resultados de nuestra investigación. Como consecuencia positiva la Universidad Ricardo Palma reduciría vacíos financieros, mejoraría la imagen y prestigio, por otro lado, los beneficios son estupendos ya que el estudiante podrá terminar su carrera y ser un profesional capacitado.

1.3. Delimitación del Proyecto

El proyecto no contempla licencias de Software.

El proyecto tiende a ser actualizado por la naturaleza de los datos de los estudiantes.

El proyecto está realizado para que en la aplicación se presione un botón y pueda cargar el flujo del análisis en general, no obstante cabe indicar que debe tener acceso a la red para que sea posible la carga de actualizaciones.

El proyecto no contempla la compra de los servidores necesarios para la realización por lo cual se emulará en ambientes virtuales.

El proyecto no asegura la velocidad de la transferencia de data ya que es soportado directamente por la Performance de los servidores.

El proyecto utilizará la tecnología Big Data para realizar un Clúster de Hadoop (HDFS) para mejorar la performance de la Base de Datos no relacional Hive.

Para la construcción del proyecto se utilizó la metodología CRISP-DM.

1.3.1. Temporal

Fecha Inicio: Enero de 2019

Fecha Fin: Mayo de 2019

1.3.2. Espacial

Universidad Ricardo Palma, Av. Benavides 5440 - Santiago de Surco
Lima.

1.3.3. Conceptual

- Machine Learning: Es una especialidad científica de Inteligencia artificial que genera sistemas las cuales puedan aprender automáticamente de acuerdo al contexto en el cual están siendo utilizados. Estos pueden identificar diversos tipos de patrones complejos por lo cual pueden generalizar comportamientos a partir de toda la información recopilada.

- **Modelo Predictivo:** Es un conjunto de técnicas estadísticas para el modelamiento, aprendizaje automatizado y minería de datos analizando todos los datos para poder predecir el futuro y generar un patrón ya que esos hechos no han sido aún registrados.
- **Deserción estudiantil:** Es la problemática del abandono de un centro de estudios, que puede ser causa de múltiples razones. Todo esto impacta demasiado y negativamente al progreso de ambos, ya que el estudiante puede no continuar los estudios y la entidad educativa generar desbalances por la disminución de los estudiantes.

1.4. Formulación del Problema

1.4.1. Problema General

- ¿De qué manera el diseño de un modelo predictivo basado en Machine Learning controla la deserción de los estudiantes en la Universidad Ricardo Palma?

1.4.2. Problemas Específicos

- ¿De qué manera el diseño de un Datamart para modelo predictivo en Machine Learning controla la deserción de los estudiantes en la Universidad Ricardo Palma?
- ¿De qué manera el diseño de los algoritmos del Machine Learning del modelo predictivo controla la deserción de los estudiantes en la Universidad Ricardo Palma?
- ¿De qué manera el diseño de una aplicación permitirá generar reportes y brindar resultados encontrados para el control de la deserción de los estudiantes en la Universidad Ricardo Palma?

1.5. Objetivos

1.5.1. Objetivo General

- Diseñar un modelo predictivo basado en Machine Learning para el control de la deserción de los estudiantes en la Universidad Ricardo Palma.

1.5.2. Objetivos Específicos

- Diseñar el Datamart del modelo predictivo en Machine Learning para el control de la deserción de los estudiantes en la Universidad Ricardo Palma.
- Diseñar los algoritmos del Machine Learning del modelo predictivo para el control de la deserción de los estudiantes en la Universidad Ricardo Palma.
- Diseñar una aplicación que permita generar reportes y brindar resultados encontrados para el control de la deserción de los estudiantes en la Universidad Ricardo Palma.

CAPÍTULO II:

MARCO TEÓRICO

2.1. Antecedentes

La idea del diseño de un modelo predictivo para disminuir la deserción de los estudiantes puede mejorar mucho la toma de decisiones oportunas por parte de las autoridades para evitar que produzca dicho problema. Muchas universidades en el transcurso de los últimos años han estado realizando similares proyectos de investigación para reducir esta tasa de deserción en sus universidades. Prueba de ello, son los numerosos casos de éxito de aplicación en diferentes tipos de organizaciones y departamentos dentro de las mismas, tanto de manera local como a nivel internacional, por a continuación, se presentará una síntesis de algunos trabajos de investigación dedicados a esta problemática universitaria.

Tomando en cuenta el ámbito nacional tenemos los siguientes antecedentes:

- “PROPUESTA DE UN MODELO DE BUSINESS INTELLIGENCE PARA IDENTIFICAR EL PERFIL DE DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD CIENTÍFICA DEL SUR”, presentado por los magister Gonzales Cam Celso, Rodríguez Domínguez César (Perú, 2017); que tiene por objetivo el proponer técnicas de inteligencia de negocios y minería de datos para realizar modelos predictivos que permitan relacionar las variables que afectan el proceso de deserción en la Universidad Científica del Sur con la información datos de la matrícula 2016-2. Este proyecto nos brindará información detallada sobre los estudiantes propensos a retirarse para la toma de decisiones de los departamentos correspondientes, por lo cual brindara recomendaciones para minimizar la deserción estudiantil enfocada a los grupos de alto riesgo.
- Gonzales y Domínguez (2017) llegan a la conclusión de que un eficiente manejo de la información a través de un Modelo de Business

Intelligence podrá disminuir la tasa del 15% de deserción que actualmente tiene dicha Universidad, también indicaron que no es concluyente una relación directa el tiempo de viaje que establecen los estudiantes para llegar a la Universidad, pero se identificó que principalmente los estudiantes más propensos a la deserción son de la carrera de Medicina Humana (49 estudiantes) en caso lleven más de 5 cursos en el ciclo.

- La implementación de la solución que indican involucra la inteligencia de negocios, es fundamental para tener un crecimiento sostenido indican, utilizando la información para alinear las estratégicas institucionales. La Universidad Científica del Sur, tiene una tasa de deserción de 15%, que origina que deje de recaudar aproximadamente S/. 8, 325,000 soles anuales, que representa 152% de las utilidades del año 2016, si desea lograr eficiencia y mejorar su rentabilidad.

- El modelo de Business Intelligence realizado en este trabajo tiene alguna similitud con nuestros modelos predictivos ya que al final nos dan supuestos de la población específica de los estudiantes que tienen a retirarse de la universidad.

- “MODELO PREDICTIVO DE DESERCIÓN UNIVERSITARIA DE LA CARRERA DE INGENIERÍA INFORMÁTICA EN LA UNIVERSIDAD RICARDO PALMA.”, presentado por el Ingeniero Gálvez Chambilla Melissa, Flores Cornejo Katherine (Perú, 2015); en el cual se proponen implementar el diseño de un modelo predictivo el cual apoye en el análisis de la deserción estudiantil de pregrado en la escuela de ingeniería informática de dicha universidad, utilizan como datos información real de la Universidad Ricardo Palma brindada por la Oficina Central de Computo para que pueda generar algoritmos y obtener las causas de la deserción universitaria de esta escuela.

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

- Gálvez y Flores (2015) llegan a la conclusión de que la solución ofrecida debe encontrar las particularidades del gran fenómeno de la deserción en la Universidad Ricardo Palma y explicar por qué los estudiantes terminan retirándose de la Universidad. Además, utilizando los modelos de Árboles de decisión y Clustering podrán predecir el a los s propensos a retirarse.
- Los últimos puntos serán claven para el análisis de la información que intentamos recopilar de toda la universidad Ricardo Palma ya que con los arboles podemos darnos una idea de la problemática actual que tiene la universidad en el área de pregrado.
- Este proyecto utiliza diferentes técnicas de minería de datos específicamente para la predicción, ellos utilizan un algoritmo en Machine Learning que son los arboles de decisión los cuales generan buenos resultados estadísticos para el alumnado analizado.
- Cabe indicar que este proyecto es un antecedente del que realizaremos ya que trabaja en la misma universidad, si bien es cierto solo toma en cuenta a la carrera de ingeniería informática, mientras nosotros hicimos el modelo predictivo global, utiliza modelos predictivos en Machine Learning muy interesantes en este caso Clustering que es uno de nuestros modelos predictivos que utilizaremos.
- “LAS CAUSAS DE LA DESERCIÓN ESTUDIANTIL DURANTE LOS PRIMEROS DOS AÑOS EN LAS ÁREAS DE CIENCIAS SOCIALES Y HUMANIDADES EN DOS UNIVERSIDADES DE GUAYAQUIL.”, presentado por el Doctor Franco Dueñas, Bernanda (Perú, 2017); donde se pretende recalcar que el enfoque cuantitativo y cualitativo son las causas principales de la deserción Universitaria en la ciudad de Guayaquil en Ecuador. Por lo cual propone algunas medidas para disminuir y contrarrestar la deserción universitaria.
- Franco (2017) llega a la conclusión de que al revisar factores influyentes en la deserción de universidades ecuatorianas, puede tomar más casos o muestras más diversas para identificar las causas de mejor manera y no tan llegando a lo subjetivo por lo cual identifica los factores familiares,

individuales y el rendimiento académico llevan a la deserción universitaria en mayoría de los estudiantes Universitarios. También analiza a un grupo de estudiantes las influencias familiares que puede tener, tales como la oposición ya sea por falta de solvencia económica o por decisión de los padres debido a la distancia que pueden estar las universidades. En cuanto al rendimiento académico del estudiante, solo unos pocos estudiantes presentaron este tipo de problemas lo cual generó una controversia debido a que en muchas investigaciones aledañas este ámbito es muy importante y determinante, debido a esto se establece como un factor secundario. Indica también que los factores de género motivan a estudiantes mujeres a desertar, sus obligaciones como madre a temprana edad y esposas no le permiten contar con el tiempo suficiente seguir estudiando.

- En este caso particular nos muestra un ámbito interpersonal del estudiante como los aspectos familiares que repercuten en el rendimiento académico y por ende en la deserción que es un punto que tomaremos en nuestra tesis, ya que las notas de los estudiantes, la puntualidad de los estudiantes es uno de nuestros factores principales para generar el modelo predictivo.

- “IMPLEMENTACIÓN DE UN MODELO PREDICTIVO BASADO EN DATA MINING Y SOPORTADO POR SAP PREDICTIVE ANALYTICS EN RETAILS”, presentado por los ingenieros Castro Porras Alexandra, Hernández Nunahuanca Juan (Perú, 2016); tiene por meta final desarrollar un modelo predictivo en empresas retail utilizando herramientas como data mining y SAP Predictive Analytics, abarcando principalmente los procesos del área de Planeamiento Comercial de dichas empresas, dicho modelo ayudará en la reducción monetarias en las empresa retail pronosticando las ventas.

- La implementación de este proyecto se basa en la evolución de SAP Predictive Analytics, los cuales ha tenido casos de éxitos encontrados por todo el mundo. El proyecto analiza y comprende la información obtenida para posteriormente configurar e implementar el modelo predictivo en la empresa retail con la data real de sus ventas, y utilizando algoritmos de predicción que la herramienta tiene.

- Debido a que ellos utilizan Predictive Analytics con SAP HANA, los procesos a realizar serán trabajados con las múltiples funcionalidades de la herramienta pues toda la información que genera los modelos automatizados se ejecuta de forma automática en memoria.
- Castro y Hernández (2016) llega a la conclusión de que al generar comparaciones de los algoritmos Triple Exponential Smoothing frente a las predicciones que se realizan en el área de planeamiento comercial de Topitop, se visualiza que se ha aumentado en la precisión de la proyección, disminuyendo el Error Porcentual Absoluto Medio (MAPE) en un 28.49%.
- El análisis realizado utiliza SAP Predictive Analytics que utiliza este proyecto ofrece grandes investigaciones en cuanto a modelos predictivos pues su integración con el lenguaje estadístico R que es el software que es el software que utilizaremos para realizar nuestras predicciones.
- “APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL EN LA EDUCACIÓN BÁSICA REGULAR EN LA REGIÓN DE LAMBAYEQUE”, presentado por el Ingeniero Piscoya Ordoñez Luis (Perú, 2016); que tiene por objetivo implementarla minería de datos como técnica y la metodología CRISP DM basado en series de tiempo para predecir la deserción escolar.
- El proyecto realizado selecciona las técnicas predictivas de minería de datos, donde utilizan varios modelos de series de tiempo y redes neuronales, dado que son la mejor forma de utilizarla por la información que fue recopilando.
- Piscoya (2016) realiza un análisis donde compara técnicas de minería de datos para al final evaluar y elegir la que mejor se adecua a su información que son los algoritmos de series de tiempo, en el proyecto que realizó utilizó técnicas ETS y redes neuronales por los criterios de selección. Donde las redes neuronales auto regresivo fue la que mejor confiabilidad resultó al analizar la información, tanto para el nivel primario y secundario con un 91% y 96% respectivamente. Indica también que la Red neuronal auto regresiva obtuvo el nivel de confianza más elevada en comparación a

ETS. En el tiempo de proceso al evaluar las técnicas se obtuvo que con el método red neuronal auto regresiva el tiempo promedio al ejecución.

- Posteriormente ellos construyeron una aplicación web para mostrar los resultados obtenidos al ejecutar los Modelos donde se extrae la data histórica de cada colegio analizado del datawarehouse.

- Este proyecto tiene similitudes ya que utilizan un datawarehouse mientras nosotros realizamos un Datamart y así poder filtrar toda la información que tiene la Universidad Ricardo Palma, para utilizar solo las tablas e información necesaria. Por otro lado con este proyecto podemos ver la perspectiva global de la deserción ya analiza a la educación básica regular la cual los índices a analizar son más globales que las que utilizamos nosotros ya que abarca a un contexto específico con características diferentes y particulares.

Tomando en cuenta el ámbito internacional tenemos los siguientes antecedentes:

- “MODELO PREDICTIVO PARA ESTIMAR LA DESERCIÓN DE LOS ESTUDIANTES EN UNA INSTITUTO DE EDUCACIÓN SUPERIOR”, presentado por el Bachiller Vásquez Jonathan (Chile, 2016); que tiene por objetivo implementar un modelo predictivo a instituciones de nivel superior utilizando Machine Learning, principalmente arboles de decisiones.

- Ellos identifican los predictores más importantes para cada semestre. Donde destacan que entre los más importantes para todos los semestres se encuentran el rendimiento en la PSU, el número de padres vivos, la evaluación de los s que realizaban a los profesores de manera semestral y el rendimiento académico universitario. Sin embargo, el conjunto total de predictores era distinto para todos los semestres, pudiéndose identificar una tendencia mientras se avanzaba en los semestres

- Vásquez (2016) indica que para los primeros cuatro semestres, los predictores relacionados con características preuniversitarias del estudiante fueron identificadas dentro de los primeros dos cuartiles de predictores. En específico, tales predictores eran el rendimiento PSU en todas sus

secciones, la cantidad de padres vivos, la evaluación que el estudiante realiza a los docentes y el nivel estudiantil de los padres de familia.

- Por lo cual se plantea las variables relacionadas con el potencial académico (rendimiento preuniversitario), antecedentes familiares y contexto educacional impactan la decisión del estudiante de permanecer en el programa. Esta relación identificada podría servir a la gestión de selección de candidatos por parte de las autoridades, como también, identificar qué estudiantes son potenciales desertores e implementar herramientas para evitar la salida temprana del estudiante durante los primeros cuatro semestres. Adicionalmente, para evitar la deserción en el segundo año, recomienda la extensión del programa de apoyo académico hasta finales del segundo año, puesto que el desempeño acumulado luego del segundo año es un predictor de mayor peso que los antecedentes preuniversitarios. Por lo cual, este proyecto enfatiza bastante en las notas del estudiante al finalizar resultados que es una determinante vital para nuestro proyecto de investigación.

- “ANÁLISIS DE LAS CAUSAS DE DESERCIÓN UNIVERSITARIA.”, presentado por el Ingeniero Dueñas Cifuentes, Mónica (Colombia, 2016); donde se pretende demostrar que una de las causas que ocasionan el abandono es la falta de trato a las causas que generan los análisis y diagnósticos de la deserción por parte de las instituciones y su desinterés ante el tema.

- Este análisis fue realizado a las universidades de la Colombia donde realizan un análisis cualitativamente las causas del abandono de cada estudiante en particular. Indica que la deserción universitaria no debe ser solo considerado un simple problema del estudiante; se recalca que el estudiante es el principal responsable de sus actos, no obstante la problemática de la deserción, no solo impacta a la vida del estudiante si no también puede generar situaciones de problemática económica y cultural. Este fenómeno seguirá de alguna forma u otra ya sea que el estudiante cambie de centro universitario. Dentro del estudio comenta las formas de cómo evitar estos altos índices de problemática. Comenta el autor que la

base de su información fue recopilada a través encuestas y conversaciones que tenía con sus entrevistados. El indica que es contexto de los estudiantes es indispensable para poder generar las causas. Dueñas (2016) indica que aun teniendo todos los indicadores es difícil establecerlos todos, no obstante este proyecto nos trae consigo las causas finales y principales de la deserción, la investigación mencionada se enfoca en tratar de demostrar las similitudes y diferencias de deserción entre los estudiantes.

- Se menciona también que todo lo relevante con lo académico aunque parezca de poca importancia influye de manera importante al estudiante durante el ciclo universitario llegando al punto de perder el impulso de continuar. La universidad debería implementar algún centro o forma donde el estudiante pueda dar la opinión que tiene sobre la universidad. La conclusión que llegan es que debido a tantas causas simultáneas hace del estudiante que tenga baja autoestima y poco ánimo de seguir en la universidad por lo cual son más propensos a la deserción.

- Dueñas (2016) llega a la conclusión de que si bien es cierto que las causas académicas son menos relevantes son estas las que determinan en la mayoría de veces en la deserción y tomando en cuenta que dichos jóvenes tienen una autoestima baja se concluye que aquellos estudiantes con las calificaciones más bajas son los más propensos a retirarse, esta información obtenida por esta tesis nos ayuda mucho en el ámbito social y en muchos casos psicológico ya que este problema tiene consigo el área interpersonal de los estudiantes donde dicha frustración obtenida por malas calificaciones dentro de la universidad genera que dichos jóvenes tengan una autoestima y pérdidas a nivel del futuro de los países ya que este talento humano es desperdiciado y perdido. En este caso las malas calificaciones como el desaprovecho de las materias es un factor que se utiliza en nuestro proyecto de investigación.

2.2. Bases Teóricas

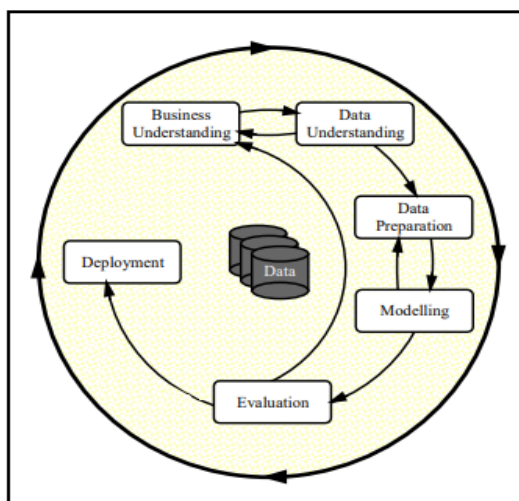
Se procederá a mostrar una serie de definiciones necesarias para poder realizar el proyecto de investigación que se realizó para encontrar las causas de deserción de la Universidad Ricardo Palma, por lo cual se muestra lo siguiente.

2.2.1 Metodología para minería de datos (CRISP-DM)

Si bien es cierto dentro de la minería de datos se tiene muchas metodologías nuestro proyecto fue realizado con la metodología CRISP-DM por la naturaleza de su arquitectura que pasaremos a explicar:

La metodología CRISP-DM para la minería de datos nos brinda una visión general del ciclo de vida de la información de un proyecto de minería de datos. Dentro de esta metodología contiene varias fases dentro del proyecto, con sus respectivas tareas y sus resultados obtenidos. Indica que el ciclo de vida de los proyectos de minería de datos se debe dividir en seis fases, que se muestran en la Figura 3. La forma de secuenciar las fases no es estricta o específica. Las flechas de la imagen indican solo las más importantes y frecuentes de las fases y las dependencias entre fases más importantes, pero en nuestro proyecto particularmente, dependería del resultado de cada fase para realizar el siguiente. (Wirth, 2016, pág. 4)

Figura 3: Niveles de la metodología CRISP-DM



Fuente: Rüdiger Wirth (2016). Phases of the Current CRISP-DM Process Model for Data Mining. [Figura2]. Recuperado de <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf>

- **Entendimiento del negocio**

Esta fase inicial se centra mayormente en comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en un problema de minería de datos y en un plan preliminar de proyecto diseñándolo para lograr los objetivos trazados. (Wirth, 2016, pág. 5)

- **Comprensión de los datos**

Esta que indica la comprensión de datos comienza con la búsqueda de los datos iniciales y seguimiento de las actividades para llegar a familiarizarse con los datos, por ello la meta es identificar problemas de calidad de los datos, el descubrimiento sobre toda información que tenga que ver con ellos, o para detectar subconjuntos interesantes para formar hipótesis ocultos dentro de los datos. Incide que existe un vínculo estrecho entre Business Understanding y Data Understanding. La formulación del problema de la minería de datos y el plan del proyecto requieren por lo menos comprensión de los datos disponibles. (Wirth, 2016, pág. 5)

- **Preparación de datos**

La fase de preparación de datos busca cubrir todas las actividades para construir el conjunto de datos final de modelado desde los datos sin realizar el procesamiento inicial. Es muy probable que las tareas de preparación de datos se realicen muchas veces y no en ningún orden prescrito. Las tareas a realizar incluirán tablas, registros y selecciones de atributos, limpieza de datos, construcción de nuevos atributos y transformación de datos para acoplarnos a las herramientas de modelado. (Wirth, 2016, pág. 5)

- **Modelado**

En esta fase, se seleccionan y aplican diversas técnicas de modelado, y sus parámetros que luego de la preparación de los datos estos son valores óptimos. Como es e conocer existe varias técnicas para la misma minería de datos tipo de problema Algunas técnicas requieren formatos de datos diferentes. (Wirth, 2016, pág. 6)

- **Evaluación**

En etapa del proyecto de investigación se ha de haber construido uno o más modelos que satisfacen nuestros objetivos ya que deben de ser de calidad, desde una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo, es importante para analizar más a fondo el modelo y ver si los pasos ejecutados son óptimos para construir el modelo, y para estar seguro de que se llega a lograr adecuadamente los objetivos de negocio. Al final de esta fase, se debe tomar una decisión sobre todo lo recopilado y el modelo elegido para el uso correcto de los resultados de la extracción de datos. (Wirth, 2016, pág. 6)

- **Despliegue del modelo**

La creación del modelo generalmente no es el final del proyecto. Por lo general, el conocimiento adquirido Tendrá que ser organizado y presentado de manera que el cliente pueda utilizarlo. Dependiendo de los requisitos, la fase de implementación puede ser tan simple como generar un

informe o tan complejo Como implementando un proceso de minería de datos repetible. En muchos casos será el usuario, no el usuario. Analista de datos, que llevará a cabo los pasos de despliegue. En cualquier caso, es importante comprender por adelantado qué acciones deberán llevarse a cabo para poder hacer uso de Los modelos creados. (Wirth, 2016, pág. 6)

2.2.2 Data Mining

En estos últimos años sobre todo desde esa necesidad que muchas empresas tienen sobre la gran revolución del Big Data se viene las constantes interrogantes de como buscar y utilizar al máximo la información que se viene almacenando por mucho tiempo con el fin de generar beneficios y oportunidades de mercado que van desde el análisis hasta poder prevenir y anticiparnos ante los requerimientos que se generan.

Data Mining o minería de datos, es el uso y la exploración de grandes cantidades de información con la finalidad de generar patrones y metas poco posibles de hallar con un sencillo análisis o modelo. También se le considera como un conjunto de procedimientos o métodos estadísticos realizados de tal forma, en algunos algoritmos de manera compleja, que nos faciliten identificar particularmente todo aquello que se aprovechar en beneficio posterior para la empresa. (Oracle, 2019, pág. 9)

2.2.3 Análisis Predictivo

Se indica que el análisis predictivo intenta descubrir todos los patrones y capturar relaciones comunes entre los datos. Las técnicas de análisis predictivo se subdividen en dos grupos. Algunas técnicas tratan encontrar los patrones históricos en la variable de resultado y utilizarlos para analizar el futuro. Otras por otro lado, como la regresión lineal, tienen como objetivo final utilizar las interdependencias entre las variables de resultado y las variables explicativas, y tratar explotarlas para realizar predicciones. Según la metodología subyacente, las técnicas pueden clasificarse en dos grupos consolidados: técnicas de regresión (por ejemplo, modelos logit multinomiales) y técnicas de aprendizaje automático (por ejemplo, redes neuronales). También existe otro tipo de clasificación que en esta ocasión se

basa en el tipo de variables de resultado: las técnicas como la regresión lineal toman las variables de resultado (por ejemplo, el precio de venta de las casas), mientras que otras, como los bosques aleatorios, se aplican a variables de resultado discretas (por ejemplo, el estado crediticio). (AmirGandomi, 2015, pág. 7)

2.2.1 Machine Learning

Es un conjunto de metodologías centradas en la estadística clásica que realizan la generación de modelos que pueden aprender de datos pasados. Este conjunto de técnicas es bastante fácil y un tanto difícil a la vez. El Aprendizaje Automático es básicamente estadística, no obstante contiene algunos detalles que diferencian de la estadística clásica regular. Se puede indicar que el Aprendizaje Automático tiene menos rigurosidad que la estadística desde un punto de vista matemático formal; el objetivo es claramente generar pruebas y demostraciones aun así está menos interesada en aportarlos y hace más énfasis en la aplicación, por lo cual utiliza la capacidad de analítica que ofrecen no solo los modernos ordenadores sino sobre todo los chips gráficos, para acelerar mucho los cálculos. Hay diferencias más sutiles, por ejemplo, en Aprendizaje Automático lo importante es la precisión del modelo construido, mientras que en estadística se hace más hincapié en cómo llevar a cabo un adecuado modelado de los datos basándose en los adecuados principios teóricos. (GIBSON, 2014, pág. 6)

2.2.2 Técnicas de Machine Learning

- **Regresión Logística:**

Una de las técnicas que ha contribuido más al avance de los modelos predictivos ha sido el avance desarrollo de determinados métodos de análisis como la regresión logística. Por lo cual se pueden hacer cuantificaciones de riesgo en un determinado carácter permitiendo al investigador la creación de modelos uni o multivariantes que sean predictivos de grandes fenómenos complejos. El modelo logístico aplicado a los estudios fue introducido por Cornfield en el año 1962 y posteriormente aplicado al análisis de los datos del estudio de Framingham. (Calvo, 2002)

- **Naive Bayes:**

Se es bien conocido que el algoritmo Naive Bayes se desempeña como bien en clasificación, no obstante su estimación de probabilidad es baja en muchas aplicaciones o modelos, por otro lado si se desea una clasificación basada en probabilidades de clase por ejemplo, una clasificación de clientes en términos de la probabilidad de que compren los productos de uno son mejores y más útiles en marketing directo. El algoritmo Naive Bayes debido a su baja estimación en el dominio binario, solo puede aprender funciones linealmente separables. Además, no puede aprender incluso todas las funciones linealmente separables. (Su, 2004, pág. 1)

- **Arboles de Decisión:**

Un árbol de decisión es una representación de una función multivariada y que su uso puede ser utilizado en diferentes ramas de la tecnología, también se es de indicar que cualquier computador con la potencia necesaria podría ejecutar el algoritmo. El interés por el uso práctico de los árboles de decisión tuvo su origen las necesidades de las ciencias sociales siendo el trabajo de SonquistMorgan (1964) el software AID (Automatic Interaction Detection). Fue uno de los primeros árboles de clasificación que se realizó para un uso práctico. Por lo cual los arboles de clasificación trascendieron, el solo ser una representación ilustrativa en los

cursos de toma de decisiones, para convertirse en una herramienta útil y sencilla de utilizar. Estos avances fueron mejorados por la obra de Breiman-Friedman-Olshen-Stone (1984) "Classification and regression trees". Un método práctico de inducción, para construir arboles de clasificación de forma recursiva, fue propuesto. Este ha sido conocido como CART. Quinlan (1986) desarrolló el algoritmo ID3 (Iterative Dichotomiser 3) este utiliza la medida de entropía de la información para crear los árboles. Esta fue mejorada y fue denominada C4.5 por su autor Quinlan (1993). Proveniente de la estadística Kass (1980) introdujo un algoritmo recursivo de clasificación no binario, llamado CHAID (Chi-square automatic interaction detection). Estos métodos permiten superar las deficiencias del AD utilizada en la Teoría Clásica de la Decisión. (Carlos N. Bouza, 2012, pág. 1)

- **Máquinas de Soporte Vectorial:**

Las máquinas de soporte vectorial son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vapnik y Cortés (1995) y su equipo, que generaron métodos relacionados con problemas de clasificación y regresión. Se han utilizado SVM para solucionar problemas de clasificación y regresión relacionados a la predicción de series de tiempo normalmente, dándonos buenos resultados en comparación a otros métodos algorítmicos. Construir las máquinas de soporte vectorial (SVM) se basa en la idea de transformar o proyectar un conjunto de datos pertenecientes a una dimensión. (Anzola, 2015, pág. 19)

2.2.3 Deserción Universitaria

Se puede decir que la deserción es la comparación cuantitativa entre la diferencia de cantidades de estudiantes que no están inscriptos comprando las cantidades de alumnado de la matrícula inicial y el fin de ciclo. Indica que existen análisis con comprobaciones científicas que normalmente son los primeros ciclos los cuales son más afectados y se localiza los estudiantes desertados. También es de considerar como deserción aquel estudiante que estudia tres meses simultáneos y al siguiente ciclo no vuelve es considerado por algunos investigadores la "primer deserción". (Peralta, 2008, pág. 66)

VARIABLES ASOCIADAS A LA DESERCIÓN

Muchas veces el origen de las causas de que un estudiante son muy complicadas de pronosticar, no obstante la deserción va directamente con variables como problemas económicos, mala elección de la carrera, falta de interés, entre otras. (Tinto, 2009, pág. 15)

En la tabla 1, se clasifican los factores de deserción por el nivel de impacto, Alta, media o baja.

Tabla 1: Factores que impactan en la deserción

Intensidad	Factores
Alta	Factores monetarios dentro de la familia. Bajo nivel de comprensión lectora y mala elección de los estudios. Mala orientación profesional.
Media	Ambientes mal adecuados dentro de la universidad Proceso educativo y acompañamiento al estudiante en su formación. Falta de programas micro curricular de investigación. Donde se debería invertir un poco ambientes más confortantes. Modelos pedagógicos universitarios desfasados o completamente diferentes a los modelos de bachillerato, que imprime un alto nivel de exigencia. Evaluaciones teóricos muy poco profundos.
Baja	Ambientes familiares. Edad del estudiante. Cantidad de oferentes. Controlar de la educación.

Fuente: Elaboración Propia

Perfil del desertor

Tomar el estudio de la deserción implica muchas características donde nos conducen a identificar las características particulares que traen con ellos.

Según (Himmel, 2002), las personas que desertan, en cualquier institución educativa, bajo cualquier circunstancia, presentan en mayor o menor grado, algunas de las siguientes características:

- Problemas de disciplina

Lo que tratan de indicarnos en este ítem es relacionado con los estudiantes que se han retirado, tal vez por impuntualidad, por temas de disciplina y en muchos casos hasta expulsados, estos son los estudiante que generan más gastos a la institución.

- Nivel socioeconómico bajo o sin opción económica

Sin duda alguna una característica que predomina como fuente común entre los desertores es el nivel económico que pasaron por lo cual tuvieron que desertar.

- Ausentismo de clases

Algunas veces por falta de dinero o por necesidad los estudiantes faltan lo que dificulta su aprendizaje y finalice con el retiro del alumno.

- Problemas de salud psicossomática

También se considera muchos casos con este tipo, donde el estudiante tiene problemas en su conducta debido problemas psiquiátricos

- Problemas inherentes a la edad

Se considera que la juventud es una de las causas de la deserción debido a que el estudiante no es lo suficientemente mayor para tomar en serio muchas veces el hecho de estudiar.

- **Inadecuadas relaciones interpersonales**

Algunos estudiantes presentan dificultades con interrelacionarse con otros compañeros lo que hace que este se aleje.

- **Resistencia a desarrollar actividades formativas**

Estos alumnos dejan de asistir a actividades extracurriculares que los podrían ayudar para mejorar sus habilidades. Tales como congresos, seminarios entre otros.

- **Inapetencia por el conocimiento**

Existen también esos estudiantes que no tienen la motivación de estudiar por lo cual solo tienden a retirarse.

- **Desmotivación hacia la carrera y/o a la universidad**

Esta es una causa muy importante ya que normalmente pasa cuando no fue seleccionada la carrera de acuerdo a su gusto o no mostro interés.

Actores que intervienen en la deserción estudiantil

Se indica en este ítem a estudiantes que se retiran por personas terceras, que los imposibilita proseguir. Se involucran, en esta casuística de la deserción, los siguientes actores:

- **Desertores**

Son los estudiantes que se han retirado por diversos motivos, en este caso universitarios.

- **Padres de Familia de los s desertores.**

Los padres llevan un papel muy importante debido a que el estudiante aún no tiene la solvencia económica en este caso para poder terminar sus estudios solos dependerá de sus padres y muchas veces por motivos monetarios viene la retirada de la Universidad, por otro lado también podría ser presión de los padres de familia por obligarle a llevar un curso diferente al que eligió.

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

Los ex compañeros pueden también persuadir a los estudiantes y querer retirarse.

- **Profesores, Directivos y administradores académicos**

Algunas veces los profesores y administrativos no hacen su trabajo como debería hacer por lo cual los estudiantes reciben malos tratos de los cuales pueden tomar malas referencias y posteriormente dar más razones para su retiro.

2.2.4 Big Data

Se le denomina Big Data a la recopilación de grandes volúmenes de datos. Los cuales con conjuntos tan grandes y complejos, conjuntos tan grandes y complejos que presentan mucha dificultad su procesamiento. Los desafíos que contemplan los requerimientos con la data son considerados como un desafío porque incluye el análisis, captura, curación, búsqueda, intercambio, almacenamiento, transferencia, visualización. Se realiza el análisis de grandes conjuntos de información de datos relacionados y se comparan para realizar subconjuntos más pequeños con la misma cantidad de datos permitiendo encontrar correlaciones para "detectar tendencias de negocios, prevenir enfermedades, combatir el crimen y así sucesivamente. Es complicado realizar trabajos con la mayoría de base de datos relacionales con grandes volúmenes de información para sistemas de gestión, estadísticas de escritorio y visualización por lo cual se requiere "software masivamente paralelo corriendo en decenas, cientos, o incluso miles de servidores " para evitar problemas. En Big Data los datos generalmente incluyen conjuntos de datos con tamaños más allá de la capacidad de herramientas de software de uso común para capturar, administrar, administrar y realizar procesamiento de datos dentro de un tiempo transcurrido tolerable ya que puede procesar de terabytes a muchos petabytes de datos. Big Data posee las 5 v que representan al volumen, velocidad, veracidad, valor y variedad. (Chavan, 2014, pág. 1)

2.2.5 Hadoop

Según (Chavan, 2014, pág. 2), Hadoop es un software de código abierto que permite la computación confiable, escalable y distribuida en clústeres de servidores baratos.

Es una implementación de código abierto de una gran escala sistema de procesamiento por lotes. Que utilizan Map-Reduce introducido por Google al aprovechar el concepto de mapa y funciones de reducción. Aunque el Hadoop framework está escrito en Java, permite a los desarrolladores desplegar programas escritos a medida codificados en Java o cualquier otro lenguaje para procesar datos de forma paralela a través de cientos o miles de servidores de productos básicos, es optimizado para solicitudes de lectura contiguas (lecturas de transmisión). Las características de Hadoop son:

- Confiable: el software es tolerante a fallos, espera y maneja fallas de hardware y software
- Escalable: diseñado para escala masiva de procesadores, memoria y almacenamiento local adjunto.
- Distribuido: Maneja la replicación, su manejo de tecnología va en programación paralelo, Map Reduce.

Sistema de archivos distribuidos de Hadoop (HDFS): un sistema de archivos distribuido que almacena datos en máquinas de productos básicos, proporcionando muy alta Ancho de banda agregado a través del clúster.

2.2.6 Metodología Scrum

Es una metodología extremadamente ágil y flexible que tiene como objetivo definir procesos de desarrollo iterativo aplicándolo a cualquier producto e inclusive en la administración y gestión de cualquier actividad que traiga consigo complejidad. proporcionando un una excelente relación entre los equipos de desarrollo. Trabaja con la participación activa de los clientes, por lo que el rendimiento del proyecto aumenta, los requisitos y la solicitud se entiende de mejor manera. Las metodologías de desarrollo ágil se vienen

destacando cada día, pero por otro lado aun no son muy distribuidas para los alumnados de las universidades. (BISSI, 2012, pág. 1)

2.3 Definición de términos básicos

- **Análisis predictivo:**

El Análisis predictivo es utilizado conjuntamente con la estadística con varios algoritmos de minería de datos para su mejor exactitud. Utilizan información histórica para realizar predicciones sobre futuros eventos

- **Árbol de decisión:**

Los arboles de decisión es una técnicas de minería de datos donde establecen un conjunto de condiciones bien organizadas de tal forma que cumplan una estructura jerárquica, de tal forma que las decisiones de la solución de la interrogante que desean solucionar puede encontrarse hasta en la última de las ramas del árbol.

- **Big Data Analytics:**

Proceso de analizar grandes volúmenes de información para descubrir patrones, sacar conclusiones y mejorar la toma de decisiones empresariales..

- **Deserción:**

Es la interrupción o el abandono de alguna actividad realizada por diferentes motivos los cuales en este caso tienen que ver con el estudio.

- **Data Mining:**

Es una de las plataformas de software para poder procesar flujos de código abierto, tiene también como objetivo final proporcionar una plataforma unificada. Utiliza varios métodos de inteligencia artificial, aprendizaje automático estadísticas, Big Data para base de datos.Su objetivo es extraer data para sintetizarla para su uso correcto.

- **Hadoop:**

Sistema de código abierto utilizado para almacenar, procesar y analizar grandes volúmenes de datos.

- **HDFS:**

Sistema de archivos distribuido, escalable y portátil que sirve como clúster de Hadoop para el alto almacenamiento de la información.

- **Deserción Universitaria**

Es la situación a la que se pueden enfrentar un estudiante cuando tiene problema alguno con proseguir adelante con los estudios.

- **Modelo Predictivo**

Es la estructura de un proceso para poder llegar a predecir utilizando un conjunto de datos y basándose en algoritmos de predicción.

- **ETL**

Proceso de extracción, transformación y carga de datos utilizando una fuente de información, base de datos o un Datamart, entre otros.

- **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.

- **Pila del sprint (Sprint Backlog)**

Tareas a realizar a través de una iteración, para construir un incremento utilizado mayormente en la metodología SCRUM.

- **Proceso:**

Es la secuencia de varios trabajos o pequeños procedimientos.

CAPÍTULO III: DESARROLLO DEL TRABAJO DE SUFICIENCIA PROFESIONAL

3.1 Modelo de Solución Propuesto

Nuestro proyecto de investigación utiliza una muestra de la población total de estudiantes de pregrado de los últimos 3 años de la Universidad Ricardo Palma, las cuales fueron analizadas para generar un Datamart donde se puede extraer la información de manera fácil, esa información recopilada será analizada por varios algoritmos en Machine Learning donde generara un análisis predictivo con la información necesaria que será exportada en una Base de Datos no relacional por lo cual se implementa un ambiente de Big Data de clúster en Hadoop para poder instalar una Base de Datos Hive, teniendo una alta performance, luego poder elaborar un sistema que genera reportes y que podrá generar realizar los procesos automáticos.

El proyecto de investigación que se realizó utiliza como base la metodología CRISP-DM la cual es perfecta para proyectos de minería de datos como es en este caso. Por lo cual se presenta lo siguiente dividido en los procesos que nos brinda la metodología CRISP-DM:

- Entendimiento del negocio

Esta fase inicial se tuvo que analizar el proyecto de una manera general o empresarial por lo cual para el mejor conocimiento y limitaciones del proyecto se tuvo reuniones con personal de la Universidad Ricardo Palma del área de OFICIC (Oficina de Centro de Computo) para poder definir lo que se busca con este proyecto, donde el objetivo común era que se elabore un diseño de un algoritmo predictivo que facilite aquellos alumnos propensos a retirarse y así tratar de tomar medidas en el asunto por parte la Universidad, también dichas reuniones se tuvo un panorama general del problema donde se entendió e identifico algunas causas globales por lo cual se daba la deserción la cual solo era intuitivas.

- **Comprensión de los datos**

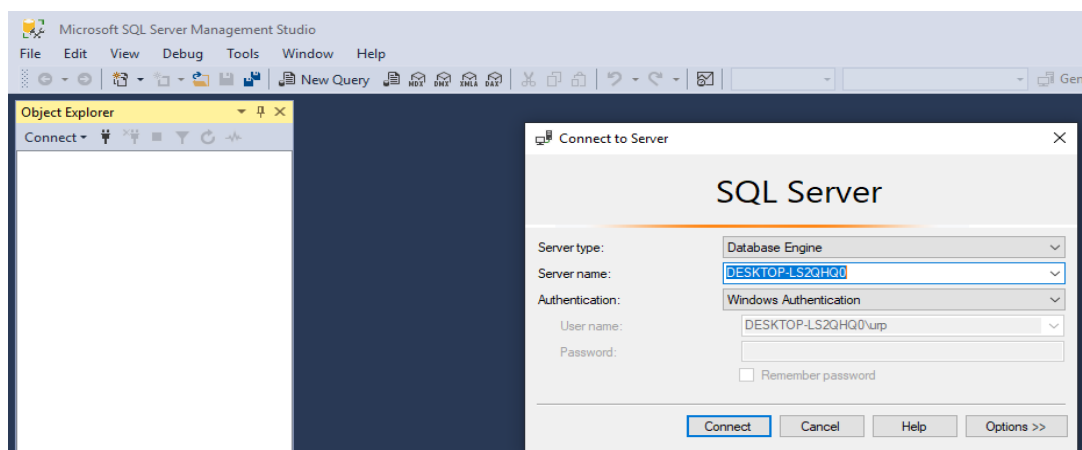
Debido a que esta fase indica el entendimiento de los datos y la búsqueda de ellas se detalla lo siguiente:

En esta parte del proyecto se procedió con los siguientes pasos para poder obtener al final el Datamart deseado.

Configuración del Software Datamart

Se tuvo que realizar el análisis de información recopilando los detalles específicos de que como está formada la Base de Datos de la URP. Por lo cual se hizo el reconocimiento de que la Universidad Trabaja con una BD Sql Server 2016 instalada sobre Windows, se observa en la figura 4 que se instaló un servidor Windows con Sql Server 2016.

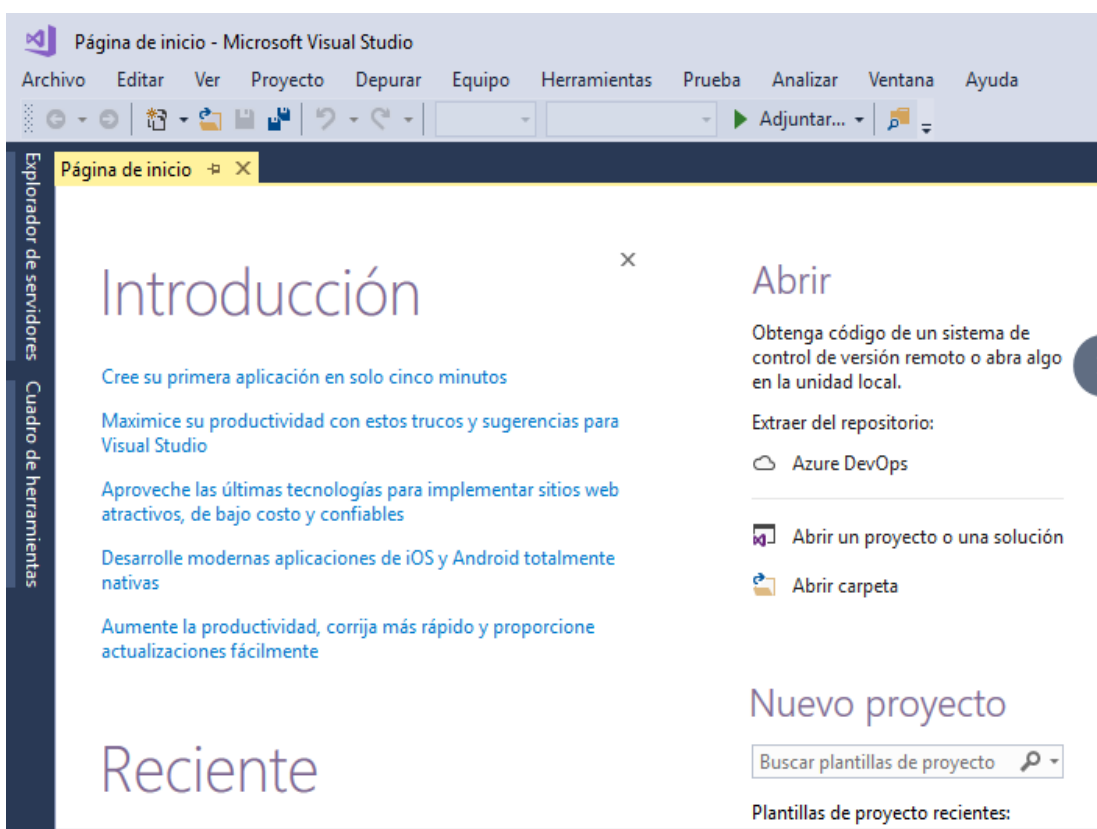
Figura 4: Instalación Sql Server 2016



Fuente: Elaboración Propia

Debido a que nuestro modelo predictivo necesita una recopilación de datos de manera automática se opta por instalar el Visual Studio para temas de preparar el Datamart, observamos la figura 5 la cual fue instalada en el mismo servidor que el Sql Server 2016 para que puedan interactuar.

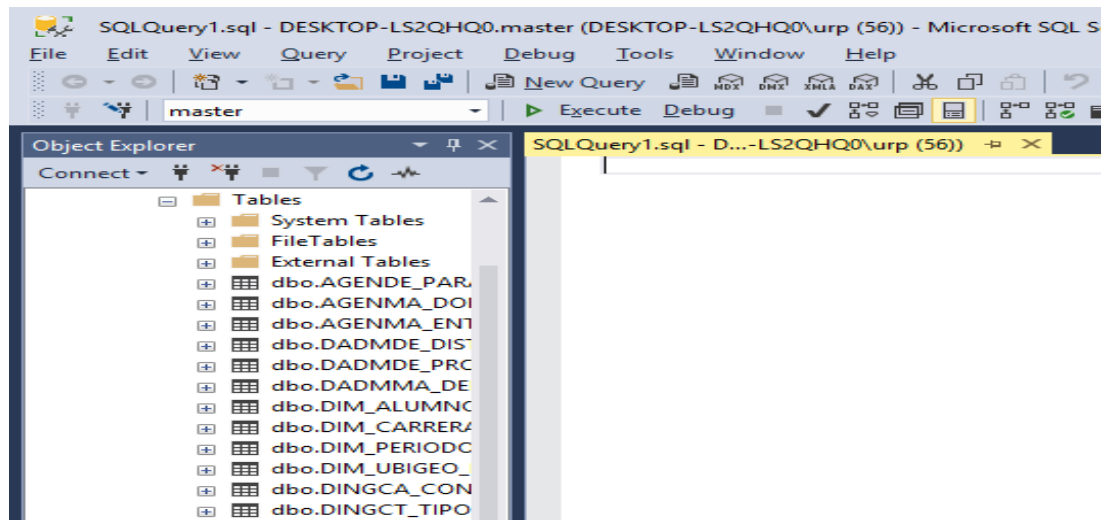
Figura 5: Instalación de Visual Studio



Fuente: Elaboración Propia

Luego de la instalación de los principales componentes a trabajar en el Datamart se da conveniente restaurar algunas tablas principales las cuales utilizaremos en nuestro servidor para realizar el Datamart, por lo cual en la figura 6 se observa que fue restaurada la Base de Datos de la URP, la cual será motivo de nuestro estudio.

Figura 6: Restauración BD URP



Fuente: *Elaboración Propia*

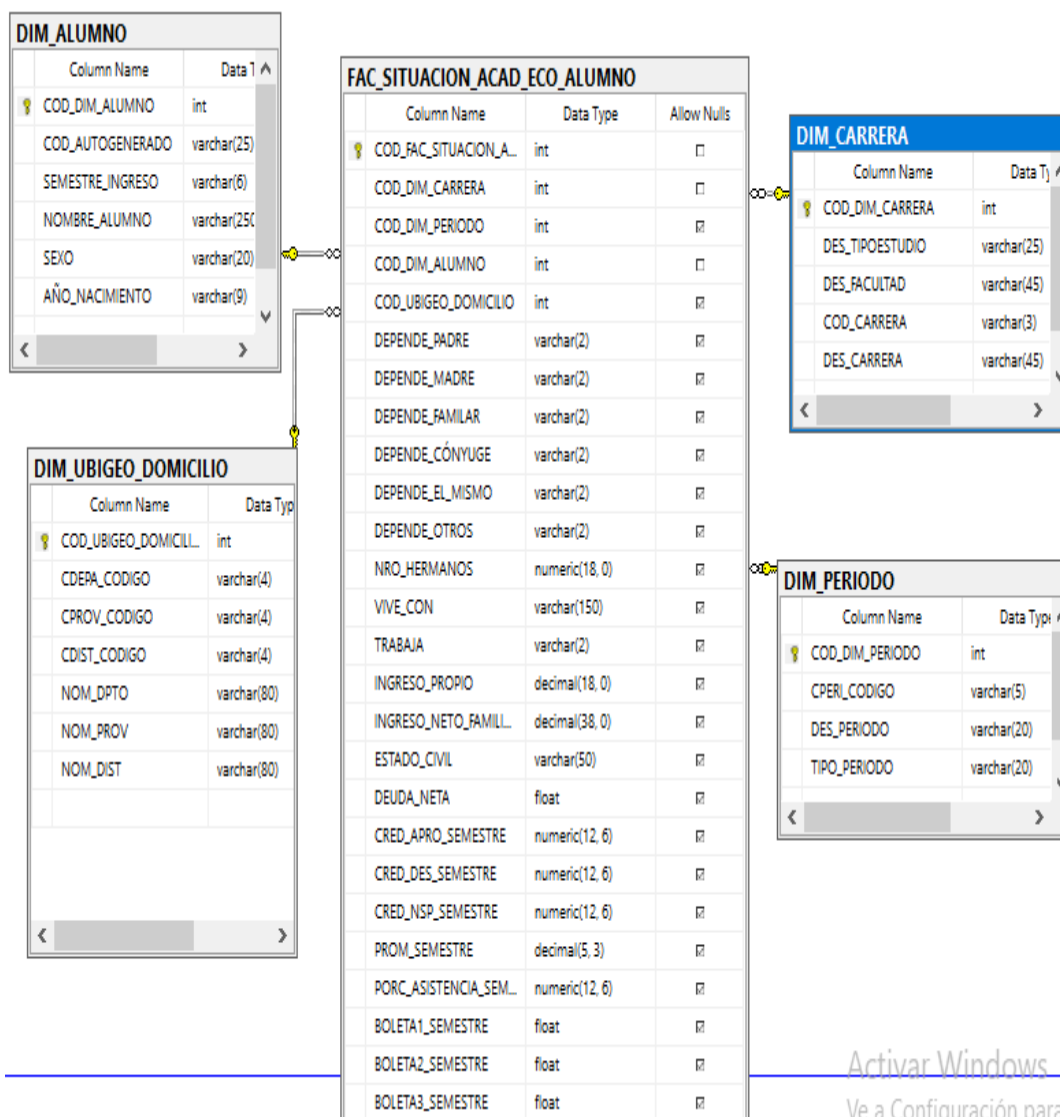
Se realizó la evaluación de la data más importante la cual posiblemente tengan influencia con los posibles de la deserción de los últimos tres años tales como:

- Promedio de notas por semestre
- Promedio deuda por semestre
- Créditos jalados por semestre
- Créditos aprobados por semestre
- Estado civil
- Deuda por semestre
- Promedio de asistencias por semestre
- Si trabaja o no
- Ingreso propio
- Ingreso neto familiar

Se ha verificado que todos estos datos son las principales causas que se podrían tomar para deserción intuitivamente. Por lo cual se registró todo datos de posible influencia.

En el grafico 7, se visualiza el esquema del Datamart.

Figura 7: Esquema del Datamart



Fuente: Elaboración Propia

Se visualiza que todos los datos de deserción están mostrados en la estructura del Datamart, por lo cual este proceso ciclo tras ciclo se hará automático debido que nuestro Datamart se alimenta de la BD de la URP.

En el gráfico 8 se muestra la creación del Datamart.

Figura 8: Creación del Datamart



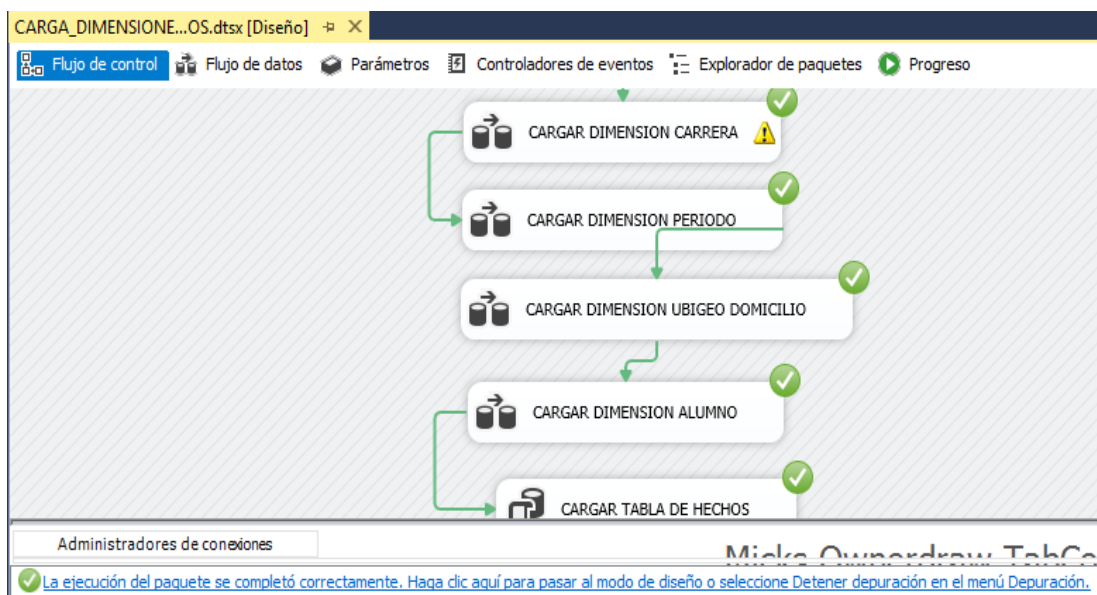
Fuente: Elaboración Propia

Se Observa que tenemos las dimensiones:

- Dimensión Limpieza: Para poder limpiar la data de todas las dimensiones y las tablas de hecho
- Dimensión de Carrera: Que tiene la información de la facultad y la carrera de los estudiantes
- Dimensión Domicilio: Que tiene solo la información del domicilio y distrito del
- Dimensión: Esta dimensión es muy importante ya que tiene información como el código, nombre, sexo.
- Tabla de Hechos: Donde recae la información general que deseamos

En la figura 9 se observa la ejecución del Datamart automático.

Figura 9: Ejecución del Datamart



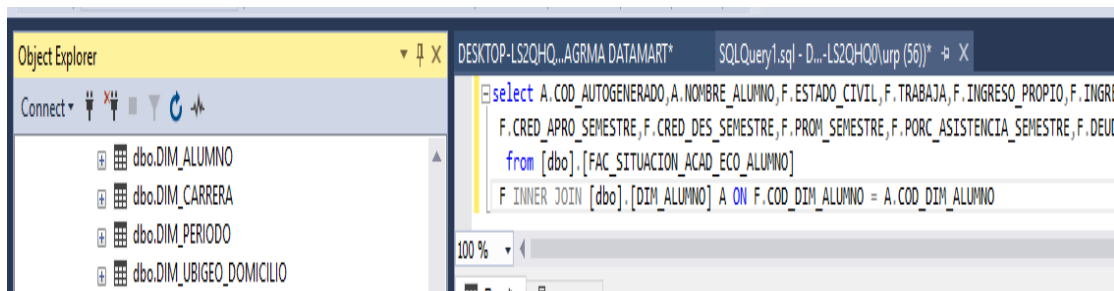
Fuente: *Elaboración Propia*

- **Preparación de datos**

En esta fase se debe tener todos los datos bien definidos para cubrir todas las actividades en este caso del análisis predictivo por lo cual se muestra lo siguientes pasos realizados para cumplirlos:

Para poder terminar con definir los datos se realiza una nueva carga en el Datamart y se hace un select al mismo para identificar solo las columnas a utilizar. Como se observa en la figura 10.

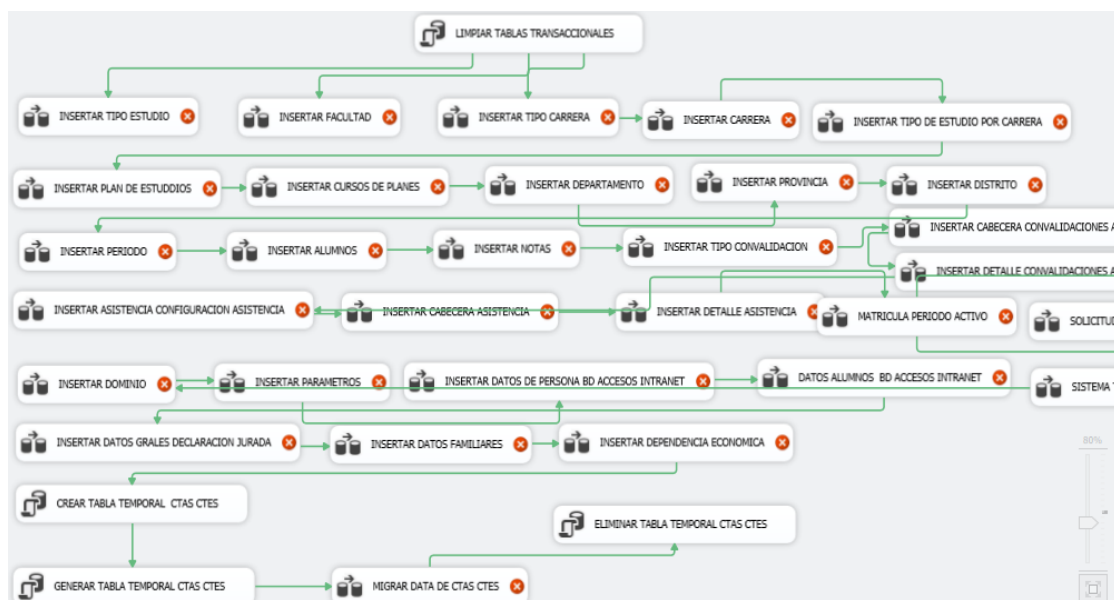
Figura 10: Nueva carga de Datamart



Fuente: Elaboración Propia

También cabe recalcar que para que este proceso sea automático cada ciclo se debe actualizar nuestra Base de Datos cada ciclo ya que solo tenemos información 2018-II. Por lo cual se creó otro Datamart para poder actualizar nuestra Base de Datos transaccional, por lo cual se muestra la figura 11.

Figura 11: Carga BD transaccional



Fuente: Elaboración Propia

Cabe recalcar que actualmente no está habilitada la opción ya que se está trabajando de manera local y no se tiene acceso a la BD Actual de la URP, no obstante si estuviera en una red Lan podría conectarse y actualizar nuestros registros.

- **Modelado**

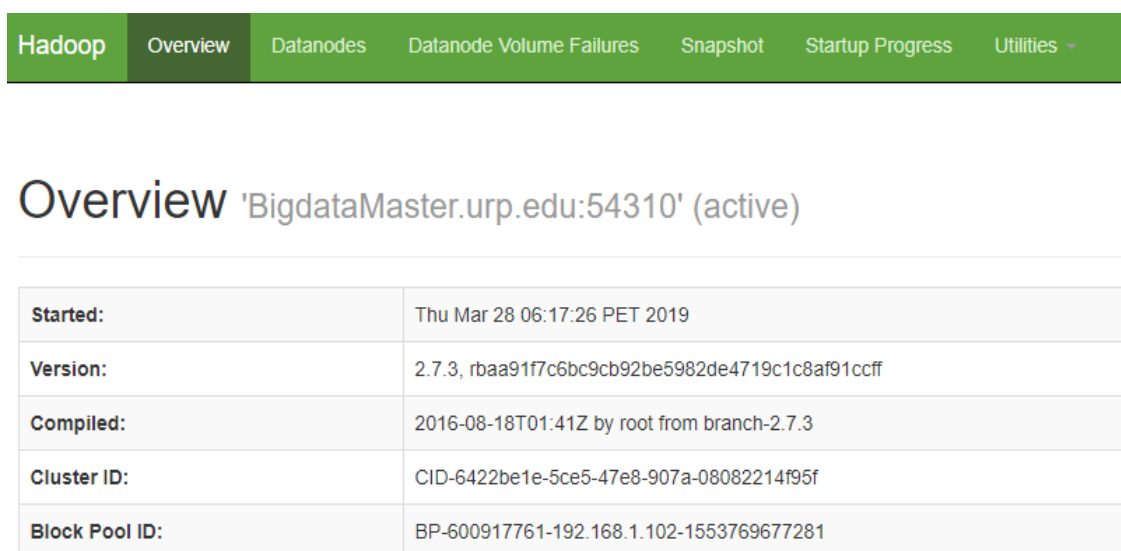
Se realiza en esta fase diversos tipos de modelado pero debido a que para el modelado se realiza se necesita configurar también el sistema operativo donde se hará el modelado. También se seleccionan y aplican diversas técnicas de modelado.

Configuración de servidores Big Data HDFS:

Para este punto en nuestro proyecto de investigación se instaló una arquitectura Big Data utilizando HDFS.

Por lo cual se muestra la figura 12, donde se muestra la interfaz del HDFS que es el clúster de Hadoop.

Figura 12: Interfaz Big Data HDFS



Fuente: Elaboración Propia

En la figura 13 se muestra más características del servidor que se ha instalado, donde se muestra un disco total de 200 gigas para labores de HDFS los cuales trabajan con un algoritmo de sistemas distribuidos para disipar la carga de los procesos. También se puede observar la versión de Hadoop instalada y el nodo del Servidor de Big Data.

Figura 13: Características de Servidor HDFS

Summary

Security is off.

Safemode is off.

3 files and directories, 1 blocks = 4 total filesystem object(s).

Heap Memory used 58.84 MB of 223 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 49.18 MB of 50.38 MB Committed Non Heap Memory. Max Non Heap Memory is -1 B.

Configured Capacity:	196.61 GB
DFS Used:	2.78 MB (0%)
Non DFS Used:	10.15 GB
DFS Remaining:	186.45 GB (94.83%)
Block Pool Used:	2.78 MB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0

Datanode Information

In operation

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
BigdataMaster.urp.edu:50010 (192.168.1.102:50010)	0	In Service	196.61 GB	2.78 MB	10.15 GB	186.45 GB	1	2.78 MB (0%)	0	2.7.3

Fuente: Elaboración Propia

En la figura 14 se muestra los procesos de HDFS (hadoop) los cuales hacen posibles poder realizar funciones y trabajos más rápidos dentro del servidor debido a su complejidad interna y sus bloques grandes la cual viene trabajando.

Figura 14: Componentes de HDFS

```
[root@bigdatamaster ~]# jps
31681 Jps
10341 NameNode
10677 SecondaryNameNode
10487 DataNode
10839 ResourceManager
11135 NodeManager
```

Fuente: Elaboración Propia

Debido a su algoritmo distribuido para poder utilizarlo se debe tener como mínimo varios nodos con un master y varios slave lo cual para efecto de pruebas se realizó solo la instalación de un servidor master, el cual también trabaja como slave por temas prácticos, no obstante para poder apreciar medio se instaló dos discos para su distribución, por lo cual se observa la figura 15 la cual se muestra el File System del servidor.

Figura 15: File System HDFS

```
[root@bigdatamaster ~]# df -h
S.ficheros      Tamaño  Usados  Disp  Uso%  Montado en
/dev/mapper/vg_root-root  50G    9,6G    41G   20%  /
devtmpfs        2,0G    0        2,0G   0%  /dev
tmpfs           2,0G    0        2,0G   0%  /dev/shm
tmpfs           2,0G    13M     2,0G   1%  /run
tmpfs           2,0G    0        2,0G   0%  /sys/fs/cgroup
/dev/sdc1       99G     63M     94G   1%  /hdfs2
/dev/sdb1       99G     66M     94G   1%  /hdfs
/dev/sda1      497M    160M    337M  33%  /boot
/dev/mapper/centos-home  48G    773M    47G   2%  /home
tmpfs           394M    0       394M   0%  /run/user/0
tmpfs           394M    12K     394M   1%  /run/user/42
```

Fuente: Elaboración Propia

Se indica de la misma forma para la siguiente figura 16, se muestra la data la cual fue obtenida del Datamart para lo cual es exportado al Servidor Big Data con los HDFS (Hadoop) para que pueda ser trabajado. Cabe indicar sé que realizo esta modalidad debido a la complejidad de la data, ya que esta tiene muchas características especiales tales como ceros, nulos, inclusive estudiantes con muy pocos para datos para su revisión por lo cual se optó por temas prácticos exportar en Excel a los servidor HDFS lo cual se muestra en la figura 17.

Figura 16: Data recompilada y Procesada del Datamart

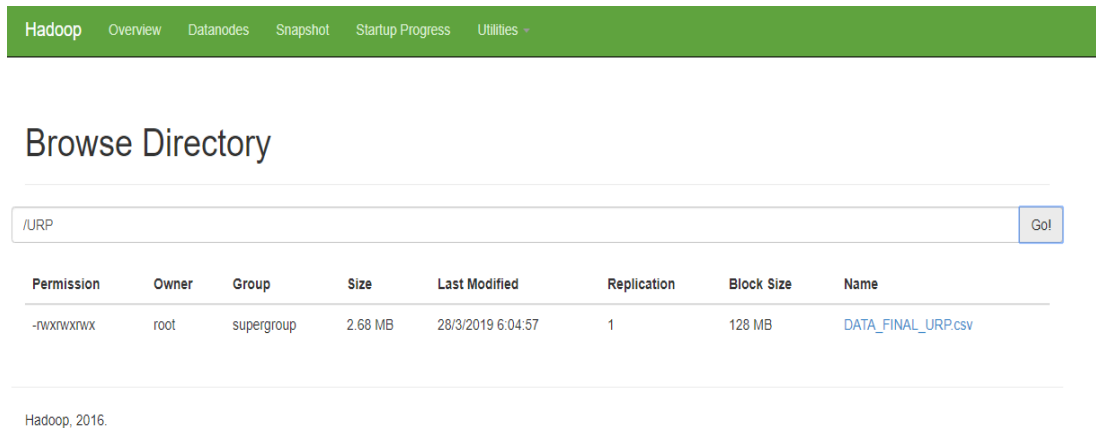
The screenshot shows a SQL query in SQL Server Enterprise Manager. The query is a SELECT statement with multiple columns and joins. The results are displayed in a table with the following data:

	SEMESTRE_INGRESO	COD_AUTOGENERADO	CPERI_CODIGO	NOMBRE_ALUMNO	ESTADO_CIVIL	TRABAJA	INGRESO_PROPI
1	20151	CAQUIDES116512015	20151	LCA/CA/DES	Soltero(a)	0	0
2	20151	ESADOSON100532015	20151	TES/ES/SON	Soltero(a)	0	0
3	20151	INOSONDA118392015	20151	HIN/IN/NDA	Soltero(a)	0	0
4	20151	EZDEZNA118842015	20151	REZ/EZ/NNA	Soltero(a)	0	0
5	20151	EZIASINA100022015	20151	REZ/EZ/INA	Soltero(a)	0	0
6	20151	ANINAAIL126222015	20151	ZAN/AN/AIL	Soltero(a)	0	0
7	20151	CAORIELA126232015	20151	ICA/CA/ELA	Soltero(a)	0	0
8	20151	IAORARTO126262015	20151	RIA/IA/RTO	Soltero(a)	0	0
9	20151	PAAYAUUEL130302015	20151	MPA/PA/UUEL	Soltero(a)	0	0
10	20151	LOENOCIA124212015	20151	LLO/LO/CIA	Soltero(a)	NULL	NULL
11	20151	LOEZACIA124222015	20151	LLO/LO/CIA	Soltero(a)	0	0
12	20151	HIHEZSHI124232015	20151	CHI/HI/SHI	NULL	NULL	NULL
13	20151	PEANAAYA125002015	20151	SPE/PE/EYA	Soltero(a)	0	0
14	20151	MADÓNSTO125102015	20151	AMA/MA/STO	Soltero(a)	0	0
15	20151	ROERADRA125132015	20151	ERO/RO/DRA	Soltero(a)	0	0
16	20151	EZNCELIN125142015	20151	NEZ/EZ/LIN	NULL	0	0
17	20151	ROHEZCAR125152015	20151	ARO/RO/CAR	Soltero(a)	0	0
18	20151	EZLESDRO125162015	20151	UEZ/EZ/DRO	Soltero(a)	0	0
19	20151	ÑANTORES125172015	20151	EÑA/ÑA/RES	Soltero(a)	0	0

Fuente: Elaboración Propia

Se muestra el Excel en el servidor Bigdata HFDS.

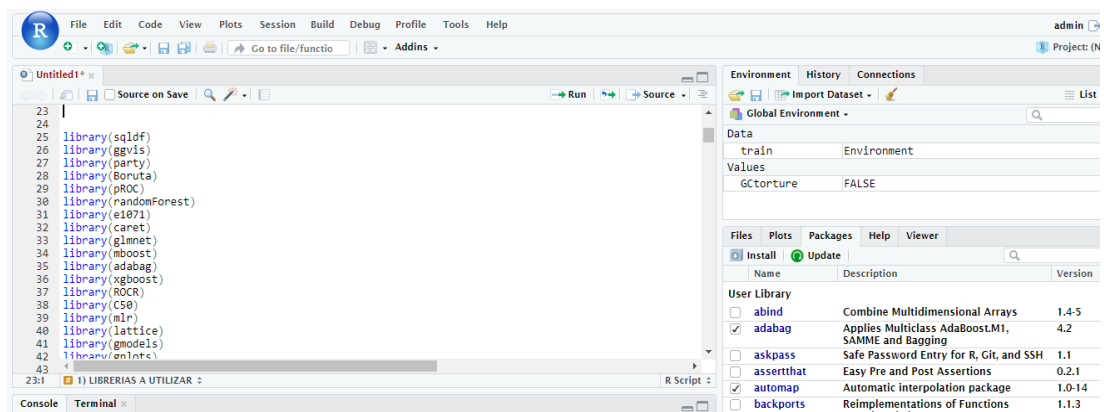
Figura 17: Excel Exportado a HDFS



Fuente: Elaboración Propia

Posteriormente en el servidor de Big Data para poder utilizarlo en el análisis predictivo se tuvo que instalar unas herramientas open source R y Rstudio, cual es óptima para realizar modelos predictivos con la data ya extraída del Datamart. Se revisa la figura 18 con dicha aplicación.

Figura 18: Aplicación Open Source Instalada R y Rstudio



Fuente: Elaboración Propia

Modelo Predictivo:

Cabe indicar que durante este proceso de Machine Learning se utilizó varios algoritmos para verificar la mejor opción a tomar. Como ya se mencionó debido a que los algoritmos aprenden de acuerdo a la data que analiza se puede optar inicialmente muchos algoritmos y conforme se realice los análisis se tomará la decisión. En las siguientes imágenes se describirá el código del modelo predictivo realizado en el software estadístico R, por lo cual en la figura 21 se visualiza todos las librerías a utilizar dentro de nuestro modelo predictivo.

Figura 19: Librerías del modelo predictivo

```
##### 1) LIBRERIAS A UTILIZAR #####  
library(rhdfs)  
library(sqldf)  
library(ggvis)  
library(party)  
library(Boruta)  
library(pROC)  
library(randomForest)  
library(e1071)  
library(caret)  
library(glmnet)  
library(mboost)  
library(adabag)  
library(xgboost)  
library(ROCR)  
library(C50)  
library(mlr)  
library(lattice)  
library(gmodels)  
library(gplots)  
library(DMwR)  
library(UBL)  
library(rminer)  
library(polycor)  
library(class)  
library(neuralnet)  
library(dplyr)  
library(MLmetrics)|  
library(data.table) # fread
```

Fuente: Elaboración Propia

Se visualiza que algunas librerías son propietarias para la utilización de algoritmos específicos. En la figura 20 se muestra la extracción de la data desde HDFS (Hadoop).

Figura 20: Extracción de Data

```
##### 2) EXTRAYENDO LA DATA #####  
  
Sys.setenv(HADOOP_CMD="/usr/local/hadoop/bin/hadoop")  
library(rhdfs)  
hdfs.init()  
train <- hdfs.file("/URP/DATA_FINAL_URP.csv","r",bufferize=104857600)  
train
```

Fuente: Elaboración Propia

En el siguiente grafico 21 se visualizara la exploración de la data ya que esta debe estar lo mejor ordenado posible. Cabe indicar que para este caso las variables inicialmente tomamos ya no estarán ya que solo se te tomará en el análisis las filas necesarias tales como:

Estado civil, si trabaja o no, Ingreso propio, ingreso neto familiar, créditos probados por semestre, créditos desaprobados, promedio de semestre, porcentaje semestral de asistencia, deuda cada semestre y el target de la deserción que en este caso sería rellenado con 1 si el estudiante deserto en el año siguiente ya que se tiene data de hace 3 años y 0 si es alumno regular o con algún crédito ese ciclo.

Figura 21: Exploración de la Data

```
68 ##### 3) EXPLORACION DE LA DATA #####
69
70 # tablas resumen
71 summary(train) # tabla comun de obtener
72 summarizeColumns(train) # tabla mas completa
73
74 resumen=data.frame(summarizeColumns(train))
75 write.csv(resumen,"resumen_data.csv",row.names=F)
76 ## Graficos para variables cuantitativas
77
```

72:1 3) EXPLORACION DE LA DATA R Script

Console Terminal x

~/Escritorio/MachineLearning/

```
> # tablas resumen
> summary(train) # tabla comun de obtener
ESTADO_CIVIL      TRABAJA      INGRESO_PROPIO      INGRESO_NETO_FAMILIAR      CRED_APRO_SEMESTRE
Length:32433      Min. :0.000      Min. : 0      Min. : 0      Min. : 0.00
Class :character  1st Qu.:0.000  1st Qu.: 0      1st Qu.: 2400      1st Qu.:10.00
Mode :character   Median :0.000  Median : 0      Median : 3800      Median :16.50
                    Mean :0.023    Mean : 673    Mean : 69194      Mean :14.94
                    3rd Qu.:0.000  3rd Qu.: 0      3rd Qu.: 6000      3rd Qu.:20.00
                    Max. :1.000    Max. :3737377  Max. :270747656    Max. :27.00
                    NA's :3600     NA's :3614     NA's :3443
CRED_DES_SEMESTRE PROM_SEMESTRE  PORC_ASISTENCIA_SEMESTRE  DEUDA_SEMESTRE  TARGETDESERCIÓN
Min. : 0.000      Min. : 0.00      Min. : 0.00      Min. : 0.0      Length:32433
1st Qu.: 0.000    1st Qu.:10.35    1st Qu.: 73.00    1st Qu.: 0.0    Class :character
Median : 3.000    Median :12.17    Median : 84.00    Median : 0.0    Mode :character
Mean : 4.286     Mean :12.12     Mean : 79.61     Mean : 321.9
3rd Qu.: 7.000    3rd Qu.:13.65    3rd Qu.: 91.80    3rd Qu.: 0.0
Max. :26.000     Max. :99.00     Max. :100.00     Max. :11226.0
NA's :150
```

Fuente: Elaboración Propia

En el grafico 22 observaremos sobre la imputación de la data, se indica que en este gráfico se analiza los factores de cada columna si el origen es cuantitativo, cualitativo, el tipo de variable numérico, integer entre otros, por lo cual la data tiene que ser imputada y convertida en numérico en su mayoría para poder realizar el trabajo de mejor manera, un claro ejemplo es la fila sexo de la data ya que normalmente cuando se tiene esta variable nos da como femenino o masculino y para efectos prácticos del modelo predictivo se convierten en 0 si es masculino y 1 si es femenino.

Figura 22: Imputación de la Data

```
78 ▾ ##### 4) IMPUTACION DE LA DATA #####
79
80 train$ESTADO_CIVIL <- as.factor(train$ESTADO_CIVIL)
81 train$y <- as.factor(train$y)
82
83 ## Imputacion Parametrica
84
85 #Podemos imputar los valores perdidos por la media o la moda
86
87 # data train
88 train_parametrica <- impute(train, classes = list(factor = imputeMode(),
89                                                    integer = imputeMode(),
90                                                    numeric = imputeMedian()),
91                               dummy.classes = c("integer", "factor"), dummy.type = "numeric")
92 train_parametrica=train_parametrica$data[,1:min(dim(train))]
93
94 summarizeColumns(train parametrica)
95 <
```

77:1 # 3) EXPLORACION DE LA DATA ↕ R Scri

Console Terminal x

~/Escritorio/MachineLearning/ ↗

```
+ dummy.classes = c( integer , factor ), dummy.type = numeric )
> train_parametrica=train_parametrica$data[,1:min(dim(train))]
> summarizeColumns(train_parametrica)
```

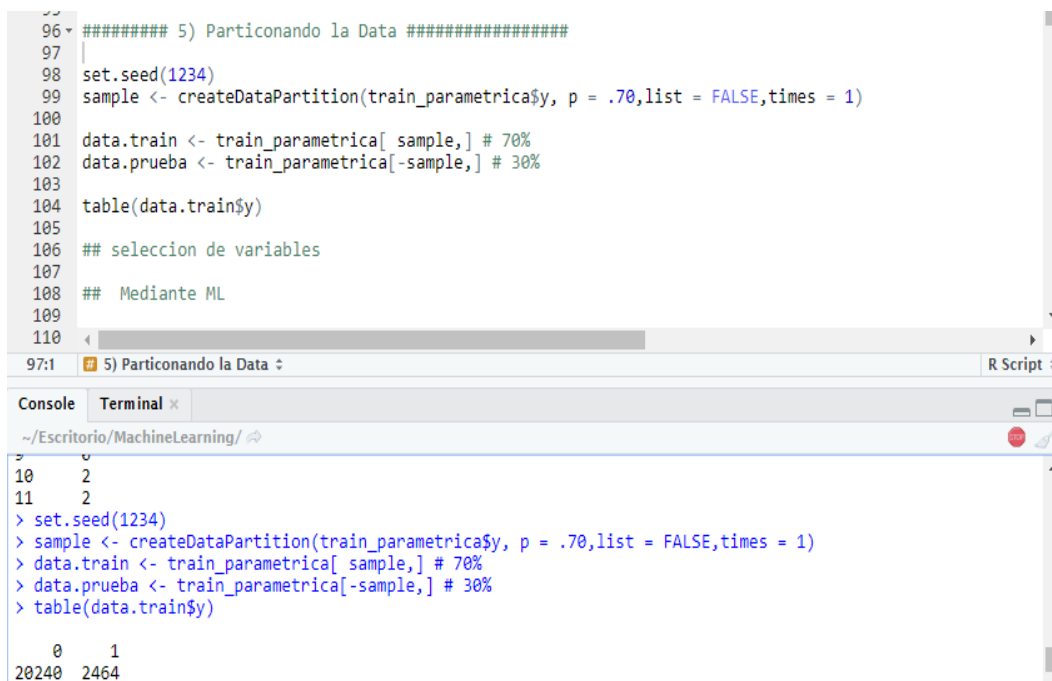
	name	type	na	mean	disp	median	mad	min	max
1	ESTADO_CIVIL	factor	0	NA	4.994913e-03	NA	NA	6	32271
2	TRABAJA	numeric	0	2.078130e-02	1.426537e-01	0.000	0.00000	0	1
3	INGRESO_PROPIO	numeric	0	5.982199e+02	4.640159e+04	0.000	0.00000	0	3737377
4	INGRESO_NETO_FAMILIAR	numeric	0	6.184844e+04	3.713828e+06	3400.000	2702.77980	0	270747656
5	CRED_APRO_SEMESTRE	numeric	0	1.493618e+01	6.796454e+00	16.500	6.67170	0	27
6	CRED_DES_SEMESTRE	numeric	0	4.285743e+00	5.161246e+00	3.000	4.44780	0	26
7	PROM_SEMESTRE	numeric	0	1.211948e+01	6.719647e+00	12.167	2.40626	0	99
8	PORC_ASISTENCIA_SEMESTRE	numeric	0	7.962988e+01	1.742134e+01	84.000	13.13160	0	100

Fuente: Elaboración Propia

A partir de esta parte es donde el algoritmo de predicción está tomando forma debido a que en la figura 23 se observará la partición de la data, este proceso es indispensable en un modelo predictivo ya que la data es partida en entrenamiento y prueba, como ya se comentó los algoritmos de Machine Learning son inteligentes por lo cual siempre una data tiene que ser probada en si misma por lo cual se tendrá 70% del total en data de entrenamiento y 30% del total en data de prueba de algoritmo.

También en este caso se muestra la utilización de una semilla la cual es utilizando por la mayoría de algoritmos, tener en cuenta que esta semilla es solo representativa pero si necesaria.

Figura 23: Particionamiento de la data



```
96 ##### 5) Particionando la Data #####
97 |
98 set.seed(1234)
99 sample <- createDataPartition(train_parametrica$y, p = .70,list = FALSE,times = 1)
100
101 data.train <- train_parametrica[ sample,] # 70%
102 data.prueba <- train_parametrica[-sample,] # 30%
103
104 table(data.train$y)
105
106 ## seleccion de variables
107
108 ## Mediante ML
109
110
```

97:1 5) Particionando la Data R Script

Console Terminal x

~/Escritorio/MachineLearning/ ↻

```
10 2
11 2
> set.seed(1234)
> sample <- createDataPartition(train_parametrica$y, p = .70,list = FALSE,times = 1)
> data.train <- train_parametrica[ sample,] # 70%
> data.prueba <- train_parametrica[-sample,] # 30%
> table(data.train$y)
 0 1
20240 2464
```

Fuente: *Elaboración Propia*

En los siguientes gráficos se mostrará los siguientes algoritmos que se analizaron por lo cual en este caso se tiene al algoritmo de Machine Learning Boruta el cual busca de manera descendente las características relevantes al comparar la significancia de los atributos originales con la importancia que se puede ganar al azar, pronostica utilizando sus copias permutadas y deshecha progresivamente las características que no tienen necesidad de usar, por lo cual se observa la figura 24.

Figura 24: Algoritmo Boruta

```
110 # Utilizando Boruta
111
112 pdf("seleccion de variables.pdf")
113 Boruta(y~.,data=data.train,doTrace=2)->Bor.hvo; # arbol cart q usa boosting
114 plot(Bor.hvo,las=3);
115 Bor.hvo$finalDecision
116 # Utilizando RF
117
118 set.seed(1234)
119 rand <- randomForest( y ~ ., data = data.train, # Datos a entrenar
120                       ntree=100,             # N?mero de ?rboles
121                       mtry = 3,              # Cantidad de variables
122                       importance = TRUE,     # Determina la importancia de las variables
123                       replace=T)            # muestras con reemplazo
124
125
126 varImpPlot(rand)
127
128 <
120:63 6) Selección de Variables ↕
```

Console Terminal x

```
~/Escritorio/MachineLearning/ ↵
DEUDA_SEMESTRAL TARGETDESECCION
Confirmed Confirmed
Levels: Tentative Confirmed Rejected
> set.seed(1234)
> rand <- randomForest( y ~ ., data = data.train, # Datos a entrenar
+                       ntree=100,             # N?mero de ?rboles
+                       mtry = 3,              # Cantidad de variables
+                       importance = TRUE,     # Determina la importancia de las variables
+                       replace=T)            # muestras con reemplazo
> varImpPlot(rand)
> library(ggvis)
```

Fuente: Elaboración Propia

El siguiente algoritmo es Naive Bayes que un clasificador de aprendizaje automatizado simple, efectivo y de uso común. Donde clasifica utilizando la regla de decisión Máximo A Posteriori en un contexto bayesiano. Por lo cual se observa la figura 25.

Figura 25: Algoritmo Naive Bayes

```
128 # Utilizando Naive Bayes
129
130 naive <- fit(y~., data=data.train, model="naiveBayes")
131 naive.imp <- Importance(naive, data=data.train)
132 impor.naive=data.frame(naive.imp$imp); rownames(impor.naive)=colnames(data.train)
133 barplot(naive.imp$imp,horiz = FALSE,names.arg = colnames(data.train),las=2)
134 impor.naive
135 dev.off()
136
137 # matriz de correlaciones no parametricas completas
138 correlaciones=hetcor(data.train, use = "pairwise.complete.obs")
139 correlaciones
140 correlaciones=correlaciones$correlations
141
142 # guardamos las correlaciones
143 write.csv(correlaciones,"correlaciones.csv")
144
145 <-
```

```
145:1 # 7) MODELADO DE LA DATA <
```

```
Console Terminal x
~/Escritorio/MachineLearning/ ↵
> naive <- fit(y~., data=data.train, model="naiveBayes")
> naive.imp <- Importance(naive, data=data.train)
Warning message:
In mean.default(data[, i]) :
  argument is not numeric or logical: returning NA
> impor.naive=data.frame(naive.imp$imp); rownames(impor.naive)=colnames(data.train)
> barplot(naive.imp$imp,horiz = FALSE,names.arg = colnames(data.train),las=2)
> impor.naive
                                naive.imp.imp
ESTADO_CIVIL                      0.002581276
TRABAJA                            0.151204308
```

```
205 # modelo 2.- Naive Bayes
206
207 modelo2=naiveBayes(y~.,data = data.train.1)
208
209 ##probabilidades
210 proba2<-predict(modelo2, newdata=data.test.1,type="raw")
211 proba2=proba2[,2]
212
213 # curva ROC
214 AUC2 <- roc(data.test.1$y, proba2)
215 auc_modelo2=AUC2$auc
216
217 # Gini
218 gini2 <- 2*(AUC2$auc) -1
219
220 # Calcular los valores predichos
221 PRED <-predict(modelo2,data.test.1,type="class")
222 |
223 # Calcular la matriz de confusi?n
224
```

Fuente: *Elaboración Propia*

Este algoritmo es la Regresión Logística que es el modelo de regresión donde el cual revisa si una variable binomial depende, o no, de otra u otras variables, clasifica los resultados entre éxito y fracaso, por cual observaremos en la figura 26 que nuestra data también es analizada por esta regresión.

Figura 26: Algoritmo de Regresión Logístico

```
166 # modelo 1.- Logistico
167
168 modelo1=glm(y~.,data=data.train.1,family = binomial(link = "logit"))
169 summary(modelo1)
170
171 proba1=predict(modelo1, newdata=data.test.1,type="response")
172
173 AUC1 <- roc(data.test.1$y, proba1)
174
175 ## calcular el AUC
176 auc_modelo1=AUC1$auc
177
178 ## calcular el GINI
179 gini1 <- 2*(AUC1$auc) -1
180
181 # Calcular los valores predichos
182 PRED <-predict(modelo1,data.test.1,type="response")
183 PRED=ifelse(PRED<=mean(proba1),0,1) # pto de corte
184 PRED=as.factor(PRED)
185 <
181:1 # 7) MODELADO DE LA DATA ↕
```

Console Terminal x

~/Escritorio/MachineLearning/ ↗

167 13343

Number of Fisher Scoring iterations: 5

```
· proba1=predict(modelo1, newdata=data.test.1,type="response")
· AUC1 <- roc(data.test.1$y, proba1)
· ## calcular el AUC
· auc_modelo1=AUC1$auc
· ## calcular el GINI
· gini1 <- 2*(AUC1$auc) -1
· library(ggvis)
```

Fuente: Elaboración Propia

El árbol Chaid es un árbol de clasificación de toma de decisiones, donde se detectara automáticamente las interacciones del Chi cuadrado, por lo cual es un algoritmo muy fuerte para predicciones ya que las predicciones se basan en combinaciones de los valores, por lo cual se muestra la figura 27 donde se hace uso de este tipo de algoritmo de Machine Learning.

Figura 27: Algoritmo Árbol Chaid

```
242 # modelo 3.- Arbol CHAID
243
244 modelo3<-ctree(y~.,data = data.train.1,
245               controls=ctree_control(mincriterion=0.95))
246
247 ##probabilidades
248 proba3=sapply(predict(modelo3, newdata=data.test.1,type="prob"),'[[',2)
249
250 # curva ROC
251 AUC3 <- roc(data.test.1$y, proba3)
252 auc_modelo3=AUC3$auc
253
254 # Gini
255 gini3 <- 2*(AUC3$auc) -1
256
257 # Calcular los valores predichos
258 PRED <-predict(modelo3, newdata=data.test.1,type="response")
259
```

Fuente: Elaboración Propia

Se muestra en la figura 28 el algoritmo árbol de clasificación Cart que se basa en la impureza de la data y selecciona nuestra información en el corte al cual conduce el mayor decrecimiento de su impureza por lo cual consigue que las sucesiones sean homogéneas en la variable Y.

Figura 28: Algoritmo Árbol Cart

```
279 # modelo 4.- Arbol CART
280
281 arbol.completo <- rpart(y~.,data = data.train.1,method="class",cp=0, minbucket=0)
282 xerr <- arbol.completo$cptable[,"xerror"] ## error de la validacion cruzada
283 minxerr <- which.min(xerr)
284 mincp <- arbol.completo$cptable[minxerr, "CP"]
285
286 modelo4 <- prune(arbol.completo,cp=mincp)
287
288 ##probabilidades
289 proba4=predict(modelo4, newdata=data.test.1,type="prob")[,2]
290
291 # curva ROC
292 AUC4 <- roc(data.test.1$y, proba4)
293 auc_modelo4=AUC4$auc
294
295 # Gini
```

Fuente: Elaboración Propia

Se muestra en la figura 29 el algoritmo de clasificación Árbol c5.0 el cual es un árbol de decisión que nos ayudara maximizar la ganancia de información de nuestra data subdivide en muchos campos la información para poder predecir.

Figura 29: Algoritmo Árbol c5.0

```
320 # modelo 5.- Arbol c5.0
321
322 modelo5 <- C5.0(y~.,data = data.train.1, trials = 55, winnow=TRUE)
323
324 ##probabilidades
325 proba5=predict(modelo5, newdata=data.test.1,type="prob")[,2]
326
327 # curva ROC
328 AUC5 <- roc(data.test.1$y, proba5)
329 auc_modelo5=AUC5$auc
330
331 # Gini
332 gini5 <- 2*(AUC5$auc) -1
333
334 # Calcular los valores predichos
335 PRED <-predict(modelo5, newdata=data.test.1,type="class")
```

Fuente: Elaboración Propia

También utilizamos el algoritmo SVM Radial en la figura 30 que representa a los puntos de muestra en el espacio separando las clases de nuestros datos mediante un hiperplano por lo cual se llama el vector soporte y se utilizó en nuestro entrenamiento de la data.

Figura 30: Algoritmo SVM Radial

```
356 # modelo 6.- SVM Radial
357
358 modelo6=svm(y~.,data = data.train.1, kernel="radial", costo=100, gamma=1, probability = TRUE)
359
360 ##probabilidades
361 proba6<-predict(modelo6, newdata=data.test.1, decision.values = TRUE, probability = TRUE)
362 proba6=attributes(proba6)$probabilities[,2]
363
364 # curva ROC
365 AUC6 <- roc(data.test.1$y, proba6)
366 auc_modelo6=AUC6$auc
367
368 # Gini
369 gini6 <- 2*(AUC6$auc) -1
370
371 # Calcular los valores predichos
372 PRED <-predict(modelo6, data.test.1,type="class")
373
```

Fuente: Elaboración Propia

En la figura 31 también observamos un algoritmo soporte vectorial SVM Linear, muy importante este algoritmo ya que con esto nuestra data tiene variedad y esta cubre dos aspectos especiales con nuestra data.

Figura 31: Algoritmo SVM Linear

```
393 # modelo 7.- SVM Linear
394
395 modelo7=svm(y~.,data = data.train.1,kernel="linear",costo=100,probability = TRUE, method='
396
397 ##probabilidades
398 proba7<-predict(modelo7, newdata=data.test.1,decision.values = TRUE, probability = TRUE)
399 proba7=attributes(proba7)$probabilities[,2]
400
401 # curva ROC
402 AUC7 <- roc(data.test.1$y, proba7)
403 auc_modelo7=AUC7$auc
404
405 # Gini
406 gini7 <- 2*(AUC7$auc) -1
407
408 # Calcular los valores predichos
409 PRED <-predict(modelo7,data.test.1,type="class")
410
```

Fuente: *Elaboración Propia*

Luego de implementar 8 tipos de algoritmos en Machine Learning podemos analizar los resultados de forma global por lo cual en nuestra investigación se hizo comparaciones del modelo Logístico, Naive Bayes, Árbol Chaid, Árbol Cart, Árbol c50, Soporte Vectorial Machine Radial, Linear, Sigmoid. Luego del análisis a cada uno de los algoritmos nos determinara el AUC que el porcentaje de la verificación del rendimiento del modelo, también nos devuelve el GINI que tiene similar comportamiento en el rendimiento, el accuracy que es la exactitud, el error del modelo y por último la sensibilidad al cambio de nuestro modelo con otra data.

En la figura 32 se observa los resultados de la curva ROC de acuerdo a los modelos ya trabajados. Cabe indicar que debido a que nosotros necesitamos determinar qué modelo nos brinda mejor exactitud, menor error y principalmente mayor sensibilidad ya que el fin de nuestro modelo es predecir cuales son los estudiantes del 2019-I que no continuaran estudiando el siguiente ciclo ya que es objeto de nuestro modelo predictivo.

Figura 32: Resultado de los Modelos

	AUC	GINI	Accuracy	ERROR	SENSIBILIDAD
Logistico	0.75	0.49	0.76	0.24	0.60
Naive_Bayes	0.74	0.47	0.87	0.13	0.37
Arbol_c50	0.77	0.53	0.92	0.08	0.27
Arbol_CART	0.70	0.40	0.91	0.09	0.26
SVM_Radial	0.64	0.27	0.91	0.09	0.21
Arbol_CHAID	0.74	0.49	0.90	0.10	0.20
SVM_sigmoid	0.60	0.20	0.83	0.17	0.17
SVM_Linear	0.51	0.01	0.89	0.11	0.00

Fuente: Elaboración Propia

- Evaluación

En etapa del proyecto de investigación ya se ha construido varios modelos que pueden satisfacer nuestros requerimiento no obstante se tiene que elegir el modelo mejor y realizar el análisis final. Se procede a describirlo:

Por lo expuesto anteriormente nuestra data es mejor modelada con la regresión Logística ya que tiene una alta sensibilidad que nos ayudará a determinar los estudiantes que no siguieran más, si bien es cierto otros algoritmos arrojan mayor porcentaje de exactitud, este algoritmo nos precisara la mejor forma de identificar a los estudiantes desertores.

En la figura 33 se muestra la elección de nuestro algoritmo ganador y realizando y realizando la ejecución del modelo sobre la data para que nos muestre el resultado final en un excel donde visualizaremos a los estudiantes que en el 2019-1 se retiraran de la Universidad Ricardo Palma, el modelo es capaz de indicarnos específicamente los alumnos que según nuestro modelo se retirarán de la Universidad en el 2019-2, cumpliendo así con nuestro modelo y generando los resultados para poder tomar las acciones necesarias fuera el caso.

Figura 33: Modelo Predictivo Final

```
# modelo ganador logistico
train_parametrica$proba <- predict(modelo1, newdata=train_parametrica,type="response")
q2 <- quantile(train_parametrica$proba , probs = c(0.25,0.5,0.75))
q2
train_parametrica$claseLogis <- ifelse(train_parametrica$proba <= q2[1],1,
                                     ifelse(train_parametrica$proba <= q2[2],2,
                                             ifelse(train_parametrica$proba <= q2[3],3,4)))
train_parametrica$Operaciones <- 1
write.csv(train_parametrica,"data_estrategia.csv",row.names = F)
```

Fuente: *Elaboración Propia*

Se visualiza en la figura que nos da un csv que será nuestro resultado del análisis correspondiente. Por el cual se generará los reportes necesarios para poder visualizar que alumno específico según nuestro algoritmo se retirará de la Universidad Ricardo Palma.

- **Despliegue del modelo**

En esta etapa de nuestro proyecto deberá ser organizado de la mejor forma automatizada posible por lo cual se mostrará la parte final que la generación del software de reportes ya automatización

Diseño de una aplicación para visualizar los modelos y generar reportes:

En este parte se mostrará la aplicación como tal ya que se mostrará los resultados del modelo predictivo, no obstante y tomando en cuenta las nuevas tendencias se optó por utilizar la metodología ágil SCRUM para poder lograr los objetivos propuestos para optimizar los tiempos en el desarrollo de las interfaces o los reportes de los resultados.

Requerimientos (Product Backlog)

Para poder dar el número de prioridad y la estimación de los requerimientos, se utilizó el modelo de la Tabla 2.

Cada requerimiento obtendrá una ponderación de baja, media, alta y muy alta, dependiendo de la lógica del negocio, es decir que requerimiento es el más importante para la Universidad y necesita ser desarrollado primero.

Tabla 2: Tabla de Prioridades

Número	Prioridad	Complejidad
1	Baja	Fácil
2	Media	Moderada
3	Alta	Complejo
4	Muy Alta	Muy Complejo

Fuente: Elaboración Propia

En la tabla 3 se muestra los requerimientos a realizar en nuestro sistema.

Tabla 3: Requerimientos del Sistema (Product Backlog)

Id Requisito	Nombre Requisito	Nro. de prioridad	Descripción	Complejidad
REQ001	Diseño de Datamart.	4	Modelamiento de la Base Datos.	4
REQ002	Generación del modelo físico y el script en el Datamart.	4	Análisis y optimización de la Base de Datos	4
REQ003	Diseñar el prototipo del interfaz del usuario login.	4	Agregar, modificar interfaz de usuario.	4
REQ004	Maquetar la interfaz de usuario login.	4	Agregar, modificar el diseño de login.	4
REQ005	Implementar lógica del proceso login.	4	Agregar, modificar, eliminar y mostrar aplicación login.	4

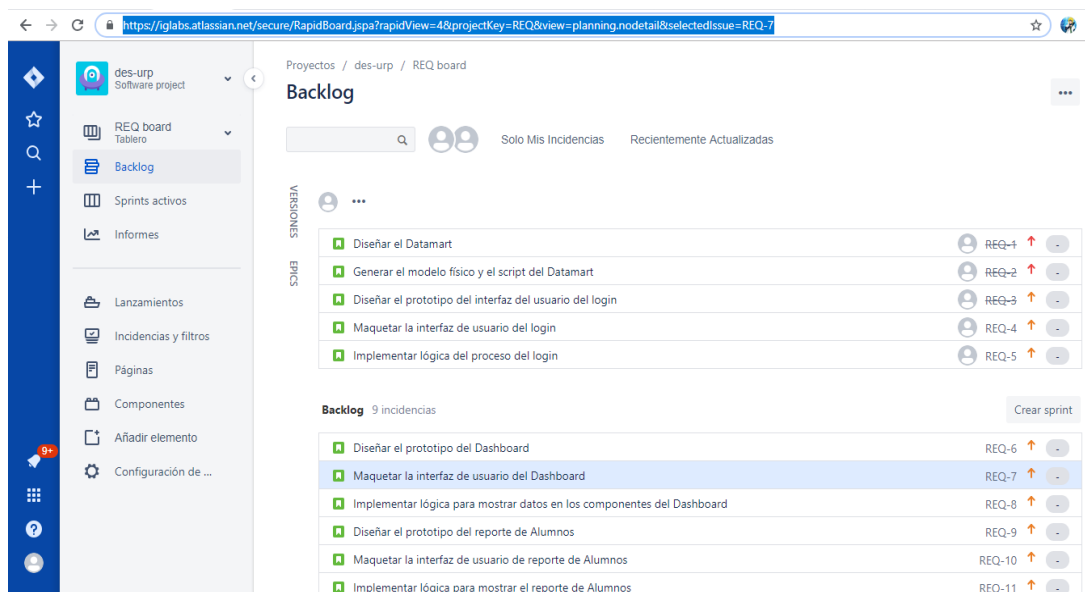
REQ006	Diseñar prototipo del Dashboard.	4	Agregar, modificar, copiar, eliminar y mostrar los prototipos de Dashboard.	3
REQ007	Maquetar la interfaz de usuario del Dashboard.	4	Agregar, modificar, copiar, eliminar y mostrar maqueta de Dashboard.	3
REQ008	Implementar lógica para mostrar datos en los componentes del Dashboard.	4	Agregar, modificar, copiar, eliminar y analizar datos de la Base de Datos	3
REQ009	Diseñar el prototipo del reporte de Alumnos.	4	Diseño de prototipo para los reportes del alumno.	3
REQ010	Maquetar la interfaz de usuario de reporte de Alumnos.	4	Agregar, modificar, copiar, eliminar y mostrar reporte de los alumnos.	3
REQ011	Implementar lógica para mostrar el reporte de Alumnos.	4	Agregar, modificar, copiar, eliminar y lógica de reportes de alumnos.	3
REQ012	Diseñar el prototipo para mostrar la carga de datos al	4	Agregar, modificar, copiar, eliminar y modelo físico de	3

	Datamart del modelo físico de la BD.		Base de Datos.	
REQ013	Maquetar la interfaz de usuario para mostrar la carga de datos del Datamart	4	Agregar, modificar, copiar, eliminar maqueta de Datamart.	3
REQ014	Implementar lógica para ejecutar la carga al Datamart del modelo físico de la Base de Datos	4	Agregar, modificar, copiar, eliminar lógica de Datamart.	3
REQ015	Diseñar la interfaz de Modelo Predictivo para mostrar la carga de datos al análisis.	4	Agregar, modificar, copiar, eliminar lógica del Modelo Predictivo.	4
REQ016	Diseñar la interfaz reportes para mostrar los resultados del análisis.	4	Agregar, modificar, copiar, eliminar lógica de resultados.	4

Fuente: *Elaboración Propia*

En la imagen 34 muestra como veníamos trabajando con una herramienta denominada Jira para el mejor control del SCRUM en los procesos de la aplicación. De la misma forma que se visualizó en la tabla anterior se observa la lista de requerimientos de nuestro proyecto DES-URP.

Figura 34: SCRUM y Jira



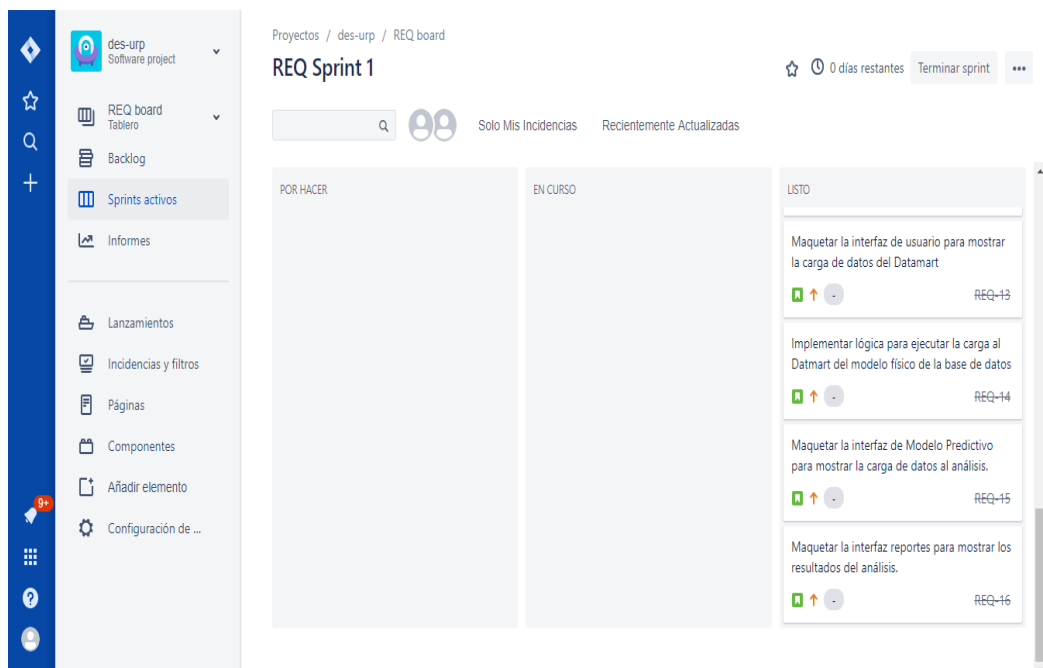
Fuente: *Elaboración Propia*

Determinación de Prioridades y Complejidad por Requerimiento:

La prioridad que se da en este caso viene ser en una sola iteración debido a que nuestro sistema no es tan complejo por lo cual por lo cual se estará realizando los requerimientos en el orden dado ya que cada uno depende del otro para la realización. Cabe indicar que para la realización del proyecto se realizó reuniones con la algunos encargados del área de sistemas por lo cual hubo tres reuniones por el sprint de nuestros requerimientos.

Se muestra la figura 35 con los requerimientos ya terminados dentro de la herramienta Jira ya comentada que trabaja en SCRUM.

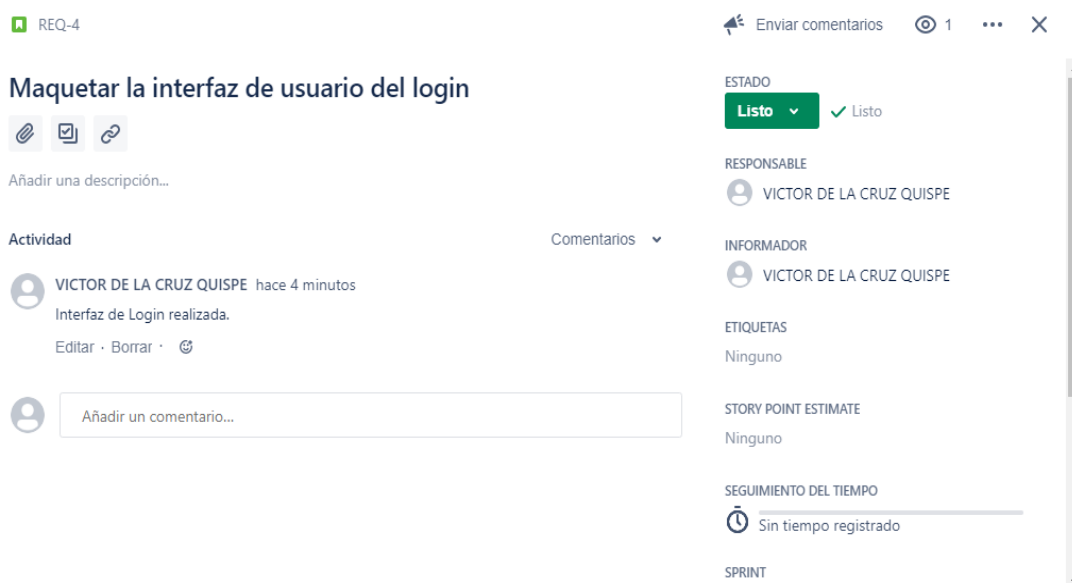
Figura 35: Requerimientos Realizados



Fuente: Elaboración Propia

En el gráfico anterior se verifica que la lista de los requerimientos pueden tener tres estados tales como por hacer, en curso y listo para un mejor control en SCRUM. Dentro de cada requerimiento se puede adjuntar alguna imagen, ingresar comentarios, designar el requerimiento, entre otros. No obstante en esta oportunidad la metodología ágil que hemos realizado para la realización de nuestra aplicación se realizó por una persona por eso en cada requerimiento se observará el nombre de la misma, se visualiza en el gráfico 36 lo indicado.

Figura 36: Diseño de los requerimientos



Fuente: Elaboración Propia

Como se viene comentando la metodología SCRUM, es utilizada en todo proceso por lo cual parte de nuestras configuraciones también fueron realizadas dentro de Jira.

- Complejidades

La asignación de complejidad para cada requerimiento se realizó como ya se había mencionado en dos reuniones con el área de Oficio de la Universidad Ricardo Palma, que es la oficina de Centro de Cómputo. Posteriormente se definió la tarea y sub tareas:

Para tareas involucradas en prototipo de interfaces tenemos:

Diseñar Interfaz

Establecer controles de interfaz

Definir estilos CSS

Establecer plantilla de aplicación con la herramienta Mockplus

Para tareas involucradas en validación tenemos:

Definir eventos

Establecer clases de la capa de datos

Establecer clases de la capa de negocio

Establecer clases de la capa de entidades

Definir objetos a partir de las clases

Por nivel Para tareas involucradas en Conceptualización:

Determinar entidades del Datamart

Determinar de atributos

Determinar de relaciones (multiplicidad)

Para tareas involucradas en Modelado

Crear de tablas independientes

Crear de tablas dependientes

Crear de relaciones

Crear de índices primarios y llaves foráneas

Análisis y Diseño (Sprint Backlog)

- Iteraciones

El sprint está comprendido por los todos requerimientos en el orden establecido donde la duración del Sprint fue de 2 meses, , se realizaron reuniones sobre algunas dudas sobre la aplicación el diseño de la data y para la presentación de los avances realizados.

En la tabla 4 se muestra las tareas realizadas de acuerdo a los requerimientos con un aproximado de hora y el responsable.

Tabla 4: Designación de tareas

Id Requisito	Nombre Requisito	Responsable	Descripción	Tiempo Estimado
REQ001	Diseño de Datamart.	Victor De la Cruz Quispe	Modelamiento de la Base Datos.	72 Horas
REQ002	Generación del modelo físico y el scipt en el Datamart.	Victor De la Cruz Quispe	Análisis y optimización de la Base de Datos	16 Horas

REQ003	Diseñar el prototipo del interfaz del usuario login.	Victor De la Cruz Quispe	Agregar, modificar interfaz de usuario.	4 Horas
REQ004	Maquetar la interfaz de usuario login.	Victor De la Cruz Quispe	Agregar, modificar el diseño de login.	4 Horas
REQ005	Implementar lógica del proceso login.	Victor De la Cruz Quispe	Agregar, modificar, eliminar y mostrar aplicación login.	4 Horas
REQ006	Diseñar prototipo del Dashboard.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar y mostrar los prototipos de Daasboard.	8 Horas
REQ007	Maquetar la interfaz de usuario del Dashboard.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar y mostrar maqueta de Dashboard.	4 Horas
REQ008	Implementar lógica para mostrar datos en los componentes del Dashboard.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar y analizar datos de la Base de Datos	4 Horas
REQ009	Diseñar el prototipo del reporte de Alumnos.	Victor De la Cruz Quispe	Diseño de prototipo para los reportes del alumno.	4 Horas

REQ010	Maquetar la interfaz de usuario de reporte de Alumnos.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar y mostrar reporte de los alumnos.	6 Horas
REQ011	Implementar lógica para mostrar el reporte de Alumnos.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar y lógica de reportes de alumnos.	3 Horas
REQ012	Diseñar el prototipo para mostrar la carga de datos al Datamart del modelo físico de la BD.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar y modelo físico de Base de Datos.	3 Horas
REQ013	Maquetar la interfaz de usuario para mostrar la carga de datos del Datamart	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar maqueta de Datamart.	4 Horas
REQ014	Implementar lógica para ejecutar la carga al Datamart del modelo físico de la Base de Datos	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar lógica de Datamart.	2 Horas
REQ015	Diseñar la interfaz de Modelo Predictivo para mostrar la carga de datos al análisis.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar lógica del Modelo Predictivo.	16 Horas
REQ016	Diseñar la interfaz reportes para mostrar los resultados del análisis.	Victor De la Cruz Quispe	Agregar, modificar, copiar, eliminar lógica de resultados.	16 Horas

Fuente: Elaboración Propia

- Criterio de determinación de prioridad y complejidad por requerimiento:
Se tomó la opinión del área de Oficio de la Universidad Ricardo Palma para los requerimientos.

Generación y seguimiento del interfaces y maquetas:

Las siguientes figuras que se tendrán son los prototipos de la aplicación y se detalla.

En la figura 37 se muestra el Prototipo del Login para el logeo de los usuarios.

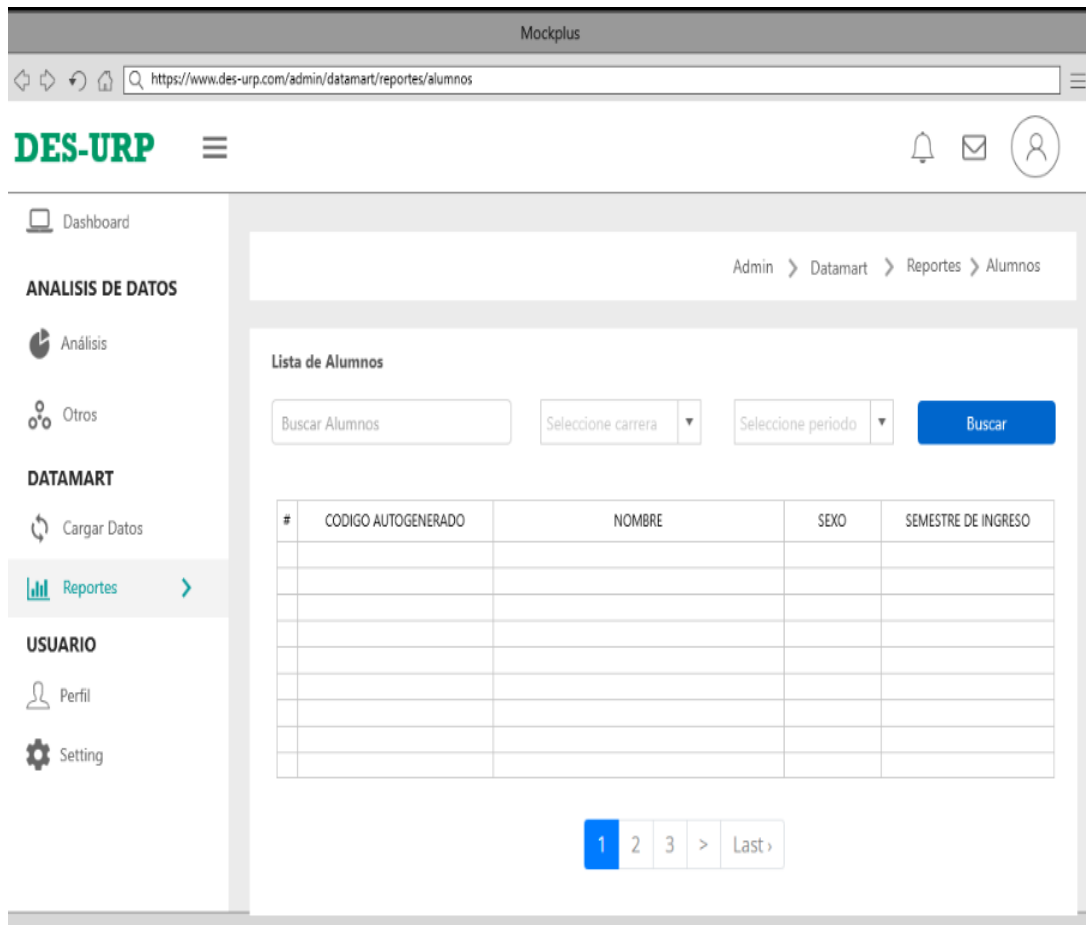
Figura 37: Prototipo Login

The image shows a login form prototype for 'DES-URP'. At the top, the text 'DES-URP' is displayed in a bold, green, sans-serif font. Below this, the form is contained within a white rectangular box with rounded corners, set against a dark gray background. The form has two input fields: the first is labeled 'Usuario' and the second is labeled 'Password'. Both labels are in a small, gray font. Below the password field is a green rectangular button with the word 'Login' written in white text.

Fuente: Elaboración Propia

En la Figura 38 se detalla el prototipo de los reportes de los estudiantes que lista a todos ellos por carrera y ciclo.

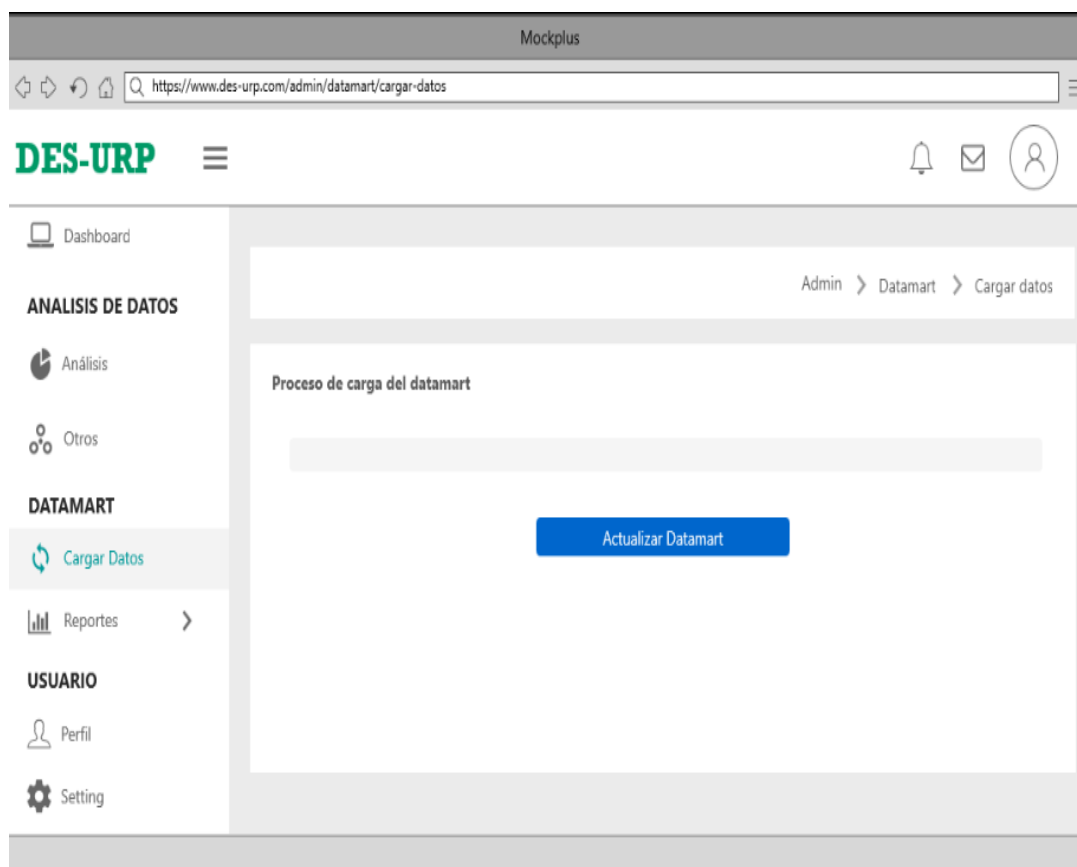
Figura 38: Prototipo de reporte estudiantes



Fuente: Elaboración Propia

En la Figura 39 se detalla el prototipo de carga de datos en el Datamart el cual se conecta a la Base de Datos del Datamart y actualiza nuevamente con la nueva información.

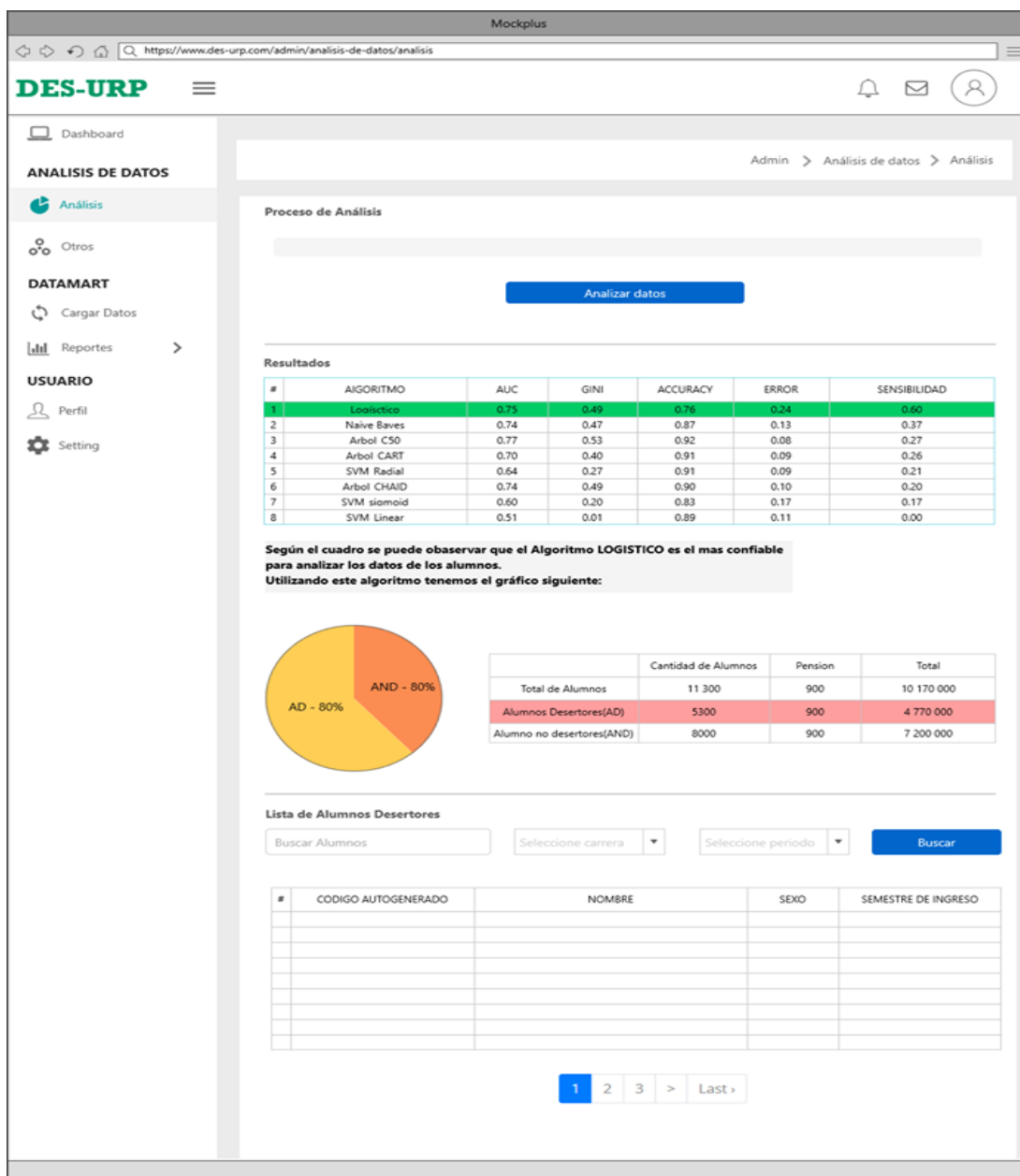
Figura 39: Prototipo carga de Datamart



Fuente: Elaboración Propia

En la Figura 40 se detalla el prototipo de Análisis predictivo el cual llama a un archivo en R en el servidor Big Data con HDFS y muestra los resultados

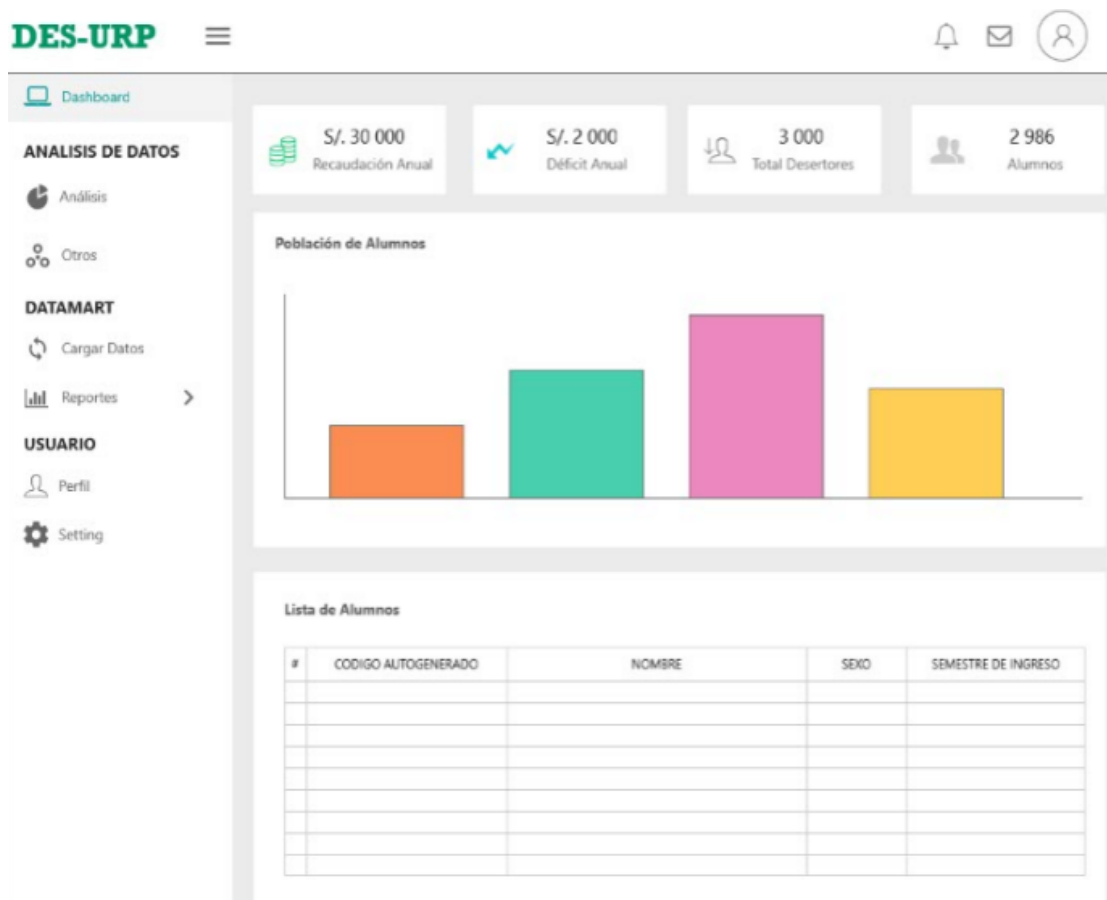
Figura 40: Prototipo Modelo Predictivo



Fuente: Elaboración Propia

En la Figura 41 se detalla el prototipo de Dashboard que nos muestra los reportes de los resultados de todo el análisis que se realizó.

Figura 41: Prototipo Dashboard

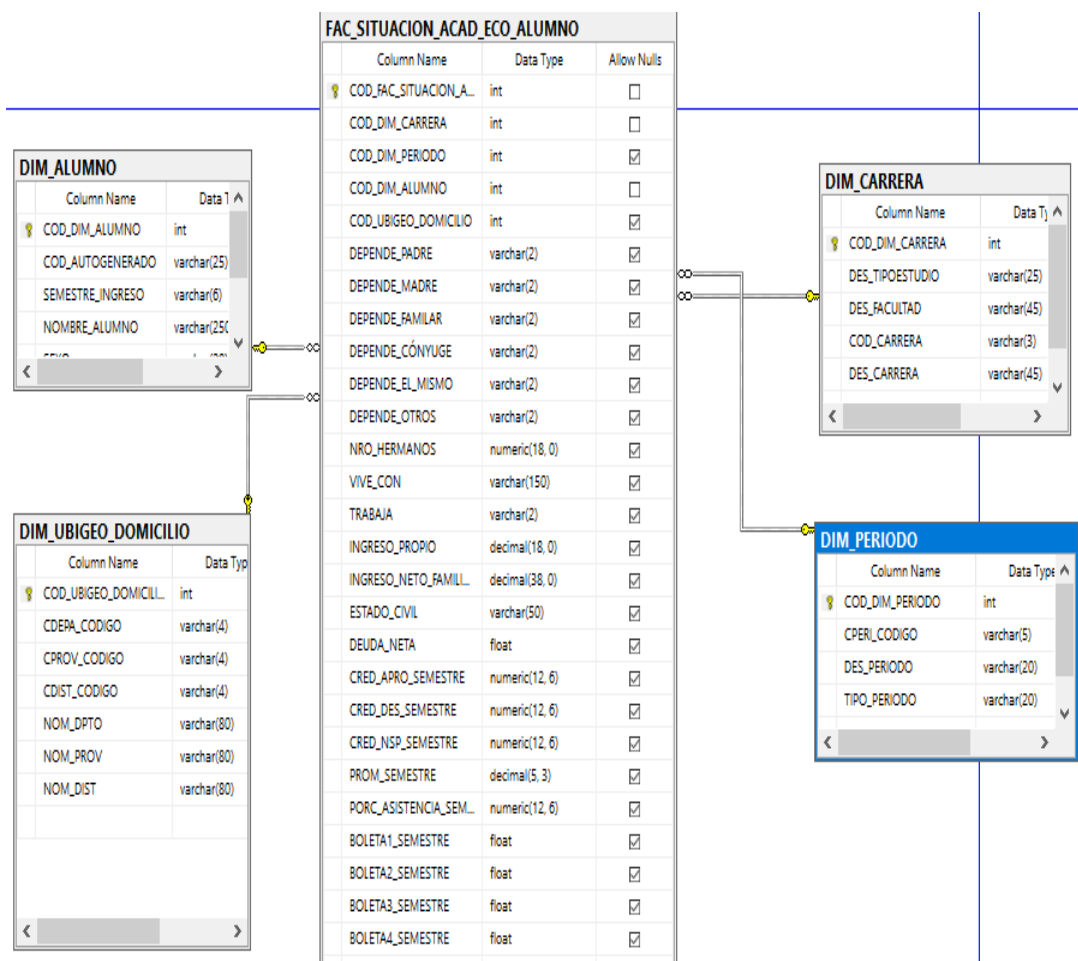


Fuente: Elaboración Propia

Evidencias sobre implementación de Software (Sprint Backlog)

A continuación mostraremos las diversas interfaces y diagramas que fueron el resultado de la aplicación de la metodología ágil SCRUM, por lo cual en el figura 42 se muestra el modelo físico de la Base de Datos que se realizó (Datamart).

Figura 42: Modelo Físico de Base de Datos Utilizada (Datamart)



Fuente: Elaboración Propia

En la figura 43 se observa el ingreso al login de nuestra aplicación que fue realizada en PHP y que registra una pequeña tabla dentro de la Base de Datos Sql Server para el almacenamiento de su información.

Figura 43: Módulo Login Aplicación

EMAIL ADDRESS

Email

PASSWORD

Password

Remember Me [Forgotten Password?](#)

SIGN IN

f SIGN IN WITH FACEBOOK

🐦 SIGN IN WITH TWITTER

Fuente: Elaboración Propia

Se observa la figura 44 donde se mostrara la interfaz de reportes de de los estudiantes tales que se puede filtrar la información por nombre del estudiantes, carrera y periodo listando en la parte derecha los alumnos identificados con la necesidad, no obstante este módulo solo contempla a todos los estudiantes sin restricciones.

Figura 44: Módulo Reporte de estudiantes

Buscar Alumno(s)

Buscar

Carrera

Periodo

Buscar

Lista de Alumnos

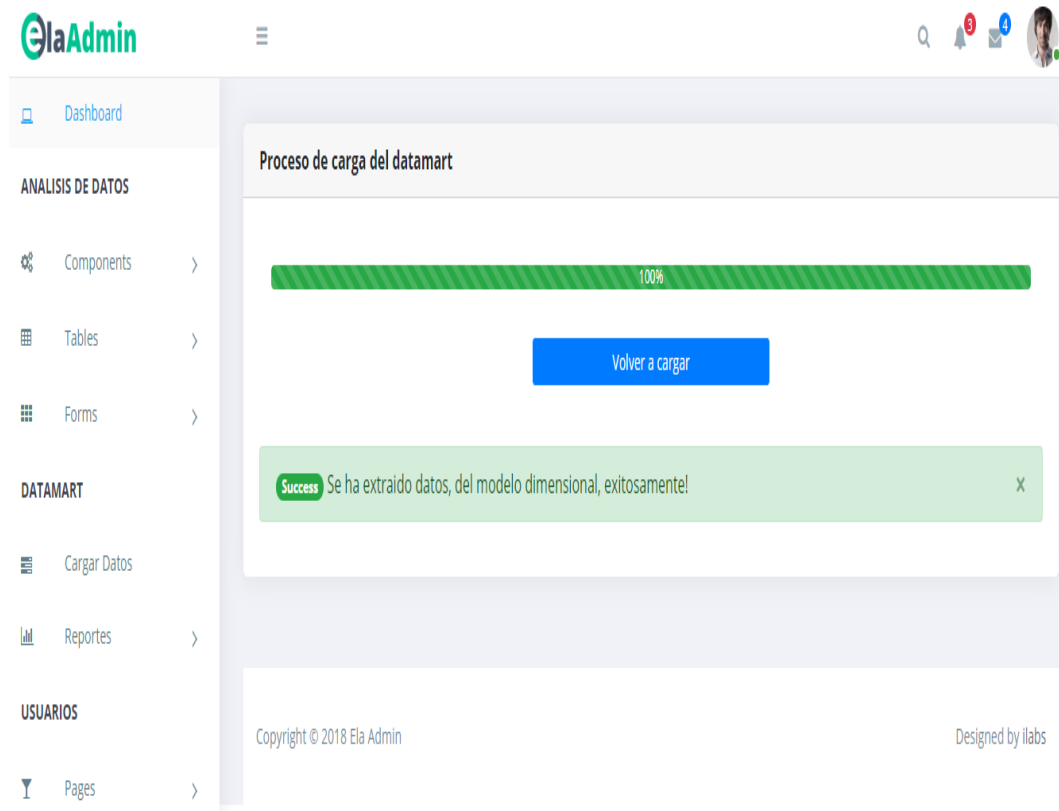
#	CÓDIGO AUTOGENERADO	NOMBRE	SEXO	SEMESTE DE INGRESO
1	IZAYAIGO210482018	UIZ/IZ/IGO	MASCULINO	20182
2	EZROAUUEL210492018	UEZ/EZ/UEL	MASCULINO	20182
3	ÁNDEZEDÚ210502018	MÁN/ÁN/EDÚ	MASCULINO	20182
4	LAQUELLA210512018	LLA/LA/LLA	FEMENINO	20182
5	ASQUELIN210522018	JAS/AS/LIN	FEMENINO	20182
6	DERIALIO210542018	RDE/DE/LIO	MASCULINO	20182
7	TOACAALO210562018	NTO/TO/ALO	MASCULINO	20182
8	ROTOSRDO210572018	RRO/RO/RDO	MASCULINO	20182
9	DOREZCIO210582018	ADO/DO/CIO	MASCULINO	20182
10	VERTOAIR210592018	AVE/VE/AIR	MASCULINO	20182

1 2 3 > Last >

Fuente: Elaboración Propia

La figura 45 nos muestra lo indicado anteriormente que es la carga del Datamart ya creado manualmente, pero debido al software se generó mediante este automatizado. Se muestra luego de presionar el botón carga de Datamart este indica el éxito de la extracción de los datos.

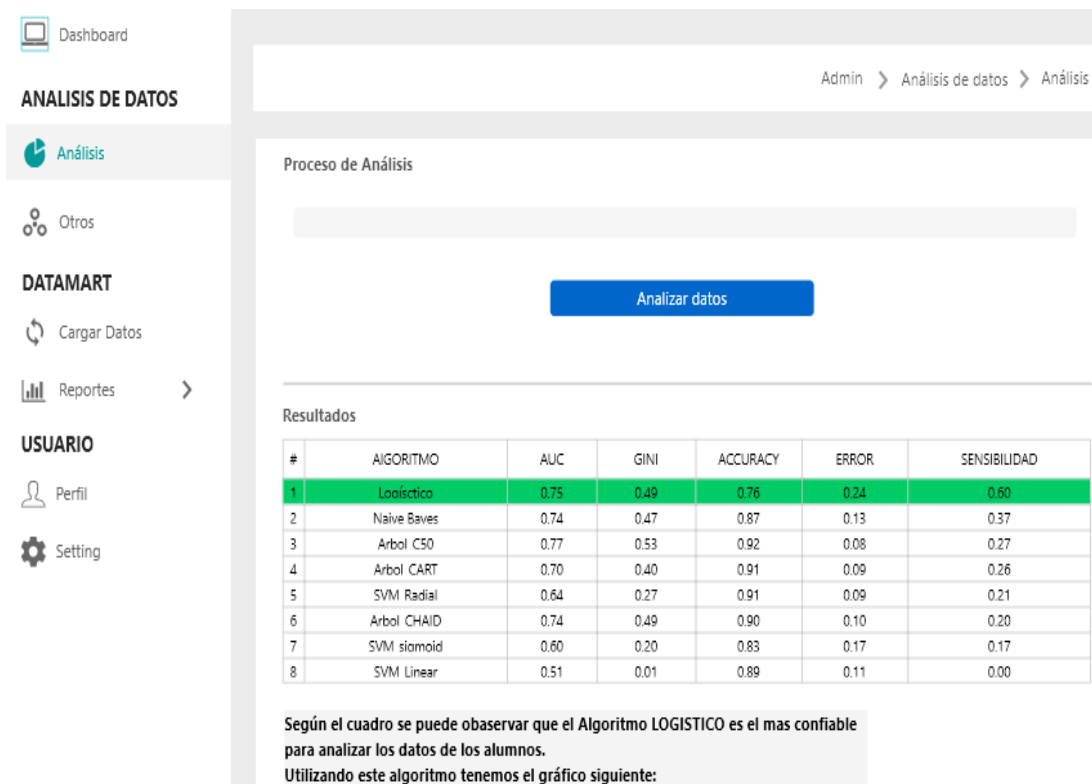
Figura 45: Módulo Carga de Datamart



Fuente: Elaboración Propia

El módulo La figura 46 nos muestra el módulo de carga de datos del Modelo Predictivo, como ya se indicó anteriormente este se conecta a un servidor Big data que tiene Instalado el R software estadístico para realización del modelo predictivo, posteriormente se exporta los excel generados a una Base de Datos y nos muestra en este caso el algoritmo ganador como se visualiza el modelo predictivo, en este caso la regresión Lineal Logística.

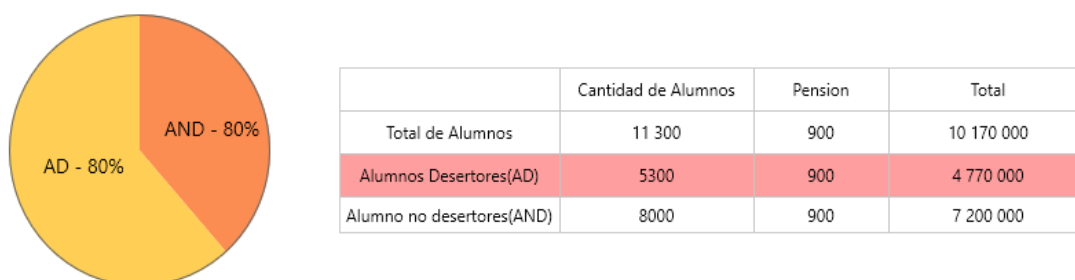
Figura 46: Módulo Carga Modelo Predictivo



Fuente: Elaboración Propia

Para finalizar en el módulo carga de datos del Modelo Predictivo está incluido el dashboard de la Deserción de los estudiantes. Donde nos indica la estadística del total de los desertores, lista los nombres de los estudiantes que según nuestro algoritmo son los desertores. Por lo cual se muestra la figura 47.

Figura 47: Listado de Desertores



Lista de Alumnos Desertores

#	CODIGO AUTOGENERADO	NOMBRE	SEXO	SEMESTRE DE INGRESO

Fuente: Elaboración Propia

Luego del despliegue de nuestro modelo predictivo en una aplicación verificamos que la metodología CRISP DM abarca todas las características que queríamos para desarrollar nuestro proyecto.

3.2 Resultados

La realización de nuestra investigación tiene como objetivo realizar un modelo predictivo capaz de identificar los principales factores de deserción e inclusive brindarnos los estudiantes posibles a desertar.

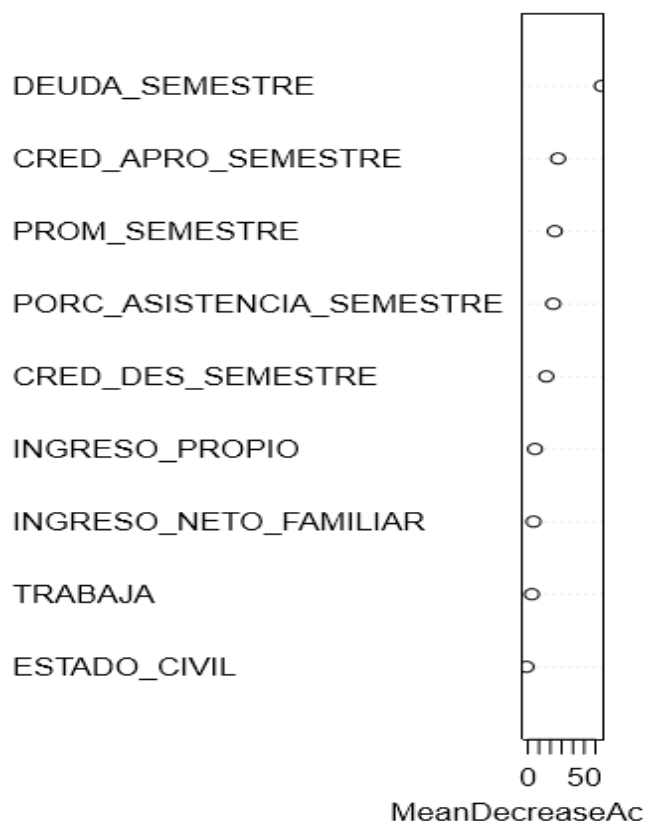
El uso de la metodología CRISP DM para la realización de nuestro proyecto benefició mucho debido a que esta metodología que utilizamos nos da las bases de los pasos a seguir para realizarlo, por lo cual hemos observado que contemplo nuestro proyecto desde la conversación de con la institución la cual es objeto de

investigación, hasta el desarrollo del software que nos brindará los resultados de nuestros modelos predictivos para evitar la deserción universitaria en la universidad Ricardo Palma.

Según el análisis con varios algoritmo de deserción se verifico que el mejor resultado del para realizar el Modelo Predictivo es con el modelo de Regresión Logístico ya que obtuvo una alta sensibilidad que nos ayudará a determinar los estudiantes que no siguieran más, si bien es cierto otros algoritmos arrojan mayor porcentaje de exactitud, este algoritmo nos precisara la mejor forma de identificar a los estudiantes desertores.

En uno de nuestros análisis se identificó que los principales factores de deserción es la deuda del semestre, créditos aprobados semestrales y promedio semestrales. Por lo cual muestro el gráfico 48 que nos trae los principales factores a tomar según el Modelo Predictivo generado en el

Figura 48: Principales factores de deserción

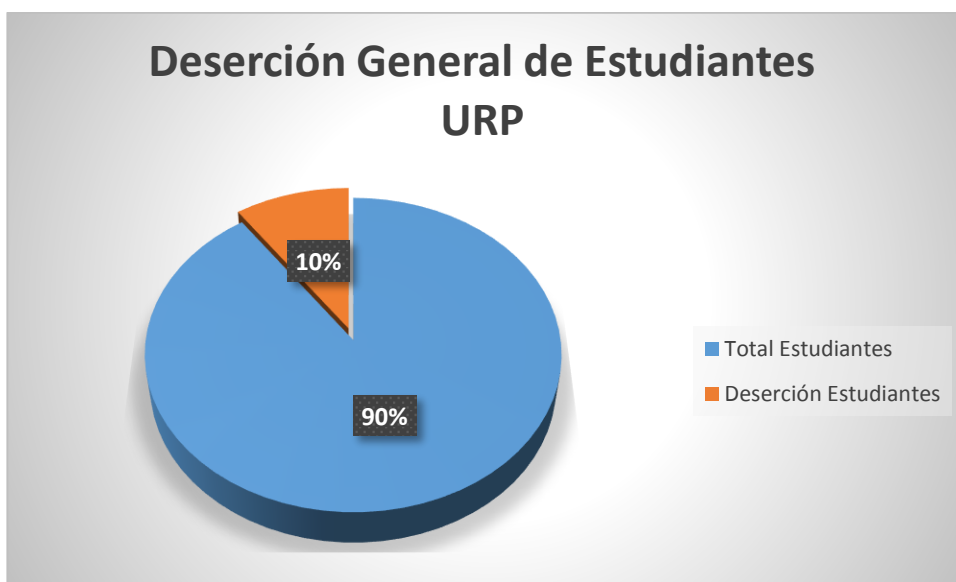


Fuente: Elaboración Propia

Se visualiza el orden de importancia de los factores que el algoritmo indica que es utilizada para realizar el modelo, por lo cual la deuda semestral que tienen los estudiantes es crucial para todo tipo de análisis, luego se muestra como los créditos aprobados semestrales la segunda opción de importancia para cualquier tipo de análisis, y en tercer lugar el promedio de asistencia semestral. Para efecto práctico la Universidad Ricardo Palma y su área de bienestar Universitario debería trabajar en esos tres aspectos principalmente para posteriormente tener mejores tasas y evitar pérdida de estudiantes y así déficit económico.

El algoritmo de predicción nos muestra nos genera los resultados e indica que la cantidad de alumnos que desertaran será el 10.85% aproximadamente del alumnado total de la Universidad Ricardo Palma, un porcentaje a muy alarmante. Por lo cual se muestra la Figura 49.

Figura 49: *Deserción General de Estudiantes URP*



Fuente: *Elaboración Propia*

Según nuestro modelo predictivo nos indica que 3520 alumnos de toda la URP desertaran de estudiar. En la siguiente tabla se mostraran los porcentajes de los estudiantes en relación a las carreras universitarias con la deserción.

Tabla 5: Porcentajes de deserción por total de desertores

Carreras Universitarias	Cuenta de DES_CARRERA	
1 Total Desertores	3520	Total
Administración de Negocios Globales	219	6.22%
Administración y Gerencia	191	5.43%
Arquitectura	522	14.83%
Biología	64	1.82%
Contabilidad y Finanzas	97	2.76%
Derecho	138	3.92%
Economía	55	1.56%
Ingeniería Civil	496	14.09%
Ingeniería Electrónica	56	1.59%
Ingeniería Industrial	381	10.82%
Ingeniería Informática	100	2.84%
Ingeniería Mecatrónica	89	2.53%
Marketing Global y Administración Comercial	102	2.90%
Medicina Humana	360	10.23%
Medicina Veterinaria	128	3.64%
Psicología	197	5.60%
Traducción e Interpretación	252	7.16%
Turismo, Hotelería y Gastronomía	73	2.07%
Total general	3520	100.00%

Fuente: Elaboración Propia

En la figura anterior se muestra los porcentajes de deserción en relación al total de los estudiantes a retirarse según el modelo. Pues se verifica que la carrera de Arquitectura e Ingeniería Civil son las más resaltantes con 14%. En la siguiente tabla se observará los porcentajes de la deserción en relación al alumnado total.

Tabla 6: Porcentaje deserción por total de Alumnos

Etiquetas de fila	Cuenta de TARGETDESERCION	DESERCIÓN	%
Administración de Negocios Globales	1762	219	12.43%
Administración y Gerencia	1765	191	10.82%
Arquitectura	5196	522	10.05%
Biología	649	64	9.86%
Contabilidad y Finanzas	923	97	10.51%
Derecho	1206	138	11.44%
Economía	452	55	12.17%
Ingeniería Civil	4746	496	10.45%
Ingeniería Electrónica	462	56	12.12%
Ingeniería Industrial	3145	381	12.11%
Ingeniería Informática	966	100	10.35%
Ingeniería Mecatrónica	816	89	10.91%
Marketing Global y Administración Comercial	771	102	13.23%
Medicina Humana	3758	360	9.58%
Medicina Veterinaria	995	128	12.86%
Psicología	2027	197	9.72%
Traducción e Interpretación	2231	252	11.30%
Turismo, Hotelería y Gastronomía	563	73	12.97%
Total general	32433	3520	10.85%

Fuente: Elaboración Propia

Se verifica que Marketing Global y Administración Comercial es la carrera con más alto índice de deserción con un 13.23% con respecto al total de sus estudiantes.

También Turismo, Hotelería y Gastronomía es otra carrera con hasta un 12.97% de alumnos que desertaran en los próximos ciclos.

CONCLUSIONES

El Luego del análisis de que se realizó se llegaron a muchas conclusiones:

Nuestro proyecto de investigación ha diseñado un modelo predictivo basado en Machine Learning utilizando el algoritmo de Regresión Logística y nos facilita el control de la deserción debido a que nos indica los las principales causas de la deserción e inclusive los estudiantes específicos que desertaran de la Universidad Ricardo Palma.

La realización del diseño de un Datamart para el modelo predictivo fue crucial ya que se utilizó la data automatizada e indispensable para poder generar el Modelo Predictivo. También el diseño nos ayudó a conocer la problemática general y los factores que son indispensables para esto, tales como la deuda semestral, los créditos aprobados semestrales, promedio de asistencias, ingreso propio, familiar entre otros.

El diseño de algoritmos en Machine Learning para realizar un modelo predictivo fue pieza clave ya que este algoritmo es ejecutado en un servidor Big Data con HDFS automatizando la performance y pudiendo probar los algoritmos: modelo Logístico, Naive Bayes, Árbol Chaid, Árbol Cart, Árbol c50, Soporte Vectorial Machine Radial, Linear, Sigmoid. Todos estos muy indispensables y con grandes expectativas, no obstante por el tema de la sensibilidad en otras datas se tomaron la solución de regresión Lineal Logístico ya tenía un 60% de estabilidad utilizándola en diferente data y cambios. Con el mismo análisis se pudo observar los porcentajes e inclusive identificar a los estudiantes con dicho perfil que según nuestro modelo dejarían de estudiar, por lo cual el cumplimiento de este objetivo fue un satisfactorio.

Para finalizar se revisó en el proyecto el diseño de una aplicación que permite generar los reportes necesarios para visualizar módulos automatizados como: el ingreso del login, la carga del datamart, la visualización de los estudiantes por carrera, carga del modelo predictivo y los reportes muy importantes para poder

visualizar los estudiantes desertores con sus carreras respectivas para poder identificarlos y tomar cartas en el asunto por parte de la URP el cual deberá generar una estrategia que será mucho más fácil debido a que se tiene los datos exactos de los estudiantes y los factores principales.

RECOMENDACIONES

Se necesita implementar el proyecto en la red de la URP para que el Datamart pueda ser alimentado con la data actual y actualizada de la universidad y no depender de la reconstrucción total de la Base de Datos, también para que las autoridades comiencen trabajar con nuestro sistema para evitar la deserción.

Se recomienda que el administrador de la aplicación constantemente revise los resultados para evitar incongruencias debido a que el sistema es automatizado no obstante no está preparado para realizar alguna acción si no tiene información con que trabajar.

También se recomienda aumentar o modificar módulos de acuerdo a nuevas necesidades que puedan tener, debido que el sistema es modular y realizado con SCRUM se puede fácilmente actualizar.

BIBLIOGRAFÍA

- 1 Wirth, R., & Hipp, J. (2016, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). Citeseer. Obtenido de (<https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf>) (Revisión: Abril de 2019)
- 2 ORACLE (2015) What is Data Mining? (http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDFGCIJ) (Revisión: Abril de 2019)
- 3 Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144. (<https://www.sciencedirect.com/science/article/pii/S0268401214001066>) (Revisión: Abril de 2019)
- 4 GIBSON, W. Posproceso estadístico 14. (http://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/Fisica_del_caos_en_la_predicc_meteo/14_Posproceso_estadistico.pdf) (Revisión: Abril de 2019)
- 5 Ortega Calvo, M., & Cayuela Domínguez, A. (2002). Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. Revista Española de Salud Pública, 76, 85-93. (<https://www.scielosp.org/pdf/resp/2002.v76n2/85-93/es>) (Revisión: Abril de 2019)
- 6 Zhang, H., & Su, J. (2004, September). Naive bayesian classifiers for ranking. In European conference on machine learning (pp. 501-512). Springer, Berlin, Heidelberg.

- (https://link.springer.com/content/pdf/10.1007%2F978-3-540-30115-8_46.pdf)(Revisión: Abril de 2019)
- 7 Bouza, C. N., & Santiago, A. (2012). La minería de datos: árboles de decisión y su aplicación en estudios médicos. *Modelación Matemática de Fenómenos del Medio Ambiente y la Salud*, 2. (https://s3.amazonaws.com/academia.edu.documents/43713947/MINERIA_DE_DATOS_MEDICOS.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1557657916&Signature=EhavbaQS3qG8COLy7IxBWDQHkg4%3D&response-content-disposition=inline%3B%20filename%3DLA_MINERIA_DE_DATOS_ARBOLES_DE_DECISION.pdf) (Revisión: Abril de 2019)
- 8 Anzola, N. S. (2016). Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. *ODEON*, (9), 113-172. (<https://revistas.uexternado.edu.co/index.php/odeon/article/view/4414/5004>) (Revisión: Abril de 2019)
- 9 Díaz Peralta, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios pedagógicos (Valdivia)*, 34(2), 65-86. (<https://www.redalyc.org/html/1735/173514136004/>) (Revisión: Abril de 2019)
- 10 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>) (Revisión Abril 2019)
- 11 Himmel, E. (2002). Modelo de análisis de la deserción estudiantil en la educación superior. *Calidad en la Educación*, (17), 91-108. (<https://www.calidadenlaeducacion.cl/index.php/rce/article/view/409>) (Revisión Abril 2019)

- 12 Chavan, V., & Phursule, R. N. (2014). Survey paper on big data. *Int. J. Comput. Sci. Inf. Technol*, 5(6), 7932-7939. (<https://pdfs.semanticscholar.org/b615/2c106010079e72fa9d5fa75f4a3172d0b033.pdf>) (Revision: Abril de 2019)
- 13 Bissi, W. (2007). Metodologia de desenvolvimento ágil. *Campo Digital*, 2(1). (Revisión: Abril de 2019)