FACULTAD DE INGENIERIA, ARQUITECTURA Y URBANISMO

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERIA DE SISTEMAS

TESIS

"APLICACIÓN DE TÉCNICAS DE MINERIA DE DATOS PARA PREDECIR LA DESERCION ESTUDIANTIL EN LA EDUCACION BASICA REGULAR EN LA REGION DE LAMBAYEQUE"

Para optar el Título Profesional de Ingeniero de Sistemas

Autor

Bach. Piscoya Ordoñez Luis Emir

ASESOR:

Ing. Carlos Alberto Chirinos Mundaca

Pimentel, 08 Marzo del 2016

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

"APLICACIÓN DE TÉCNICAS DE MINERIA DE DATOS PARA PREDECIR LA DESERCION ESTUDIANTIL EN LA EDUCACION BASICA REGULAR EN LA REGION DE LAMBAYEQUE"

Aprobación de la Tesis

Ing. Denny John Fuentes Adrianzén
Presidente del jurado de tesis

Ing. Carlos Alberto Chirinos Mundaca
Asesor especialista

Ing. Rosa América Cobeñas Sánchez
Secretario del jurado de tesis

Mg. Victor Tuesta Monteza
Vocal del jurado de tesis

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

DEDICATORIA

A Dios, por su amor y fortaleza para lograr mis metas; a mis padres, por su buen ejemplo que me hizo mejor persona; a mis maestros, que a lo largo de nuestra vida académica han iluminado mi mente y ampliado mis conocimientos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

AGRADECIMIENTO

A mi familia, por alentarme a seguir adelante, aún en los momentos más difíciles. También mi reconocimiento a las personas que han hecho posible la culminación de este trabajo de investigación; al Mg. Carlos Alberto Chirinos Mundaca quien fue mi Asesor Especialista de mi Tesis, el cual fue una pieza fundamental para la realización de este proyecto, quien con sus sabios consejos, orientaciones, revisión y corrección, permite la conclusión correcta de este documento; también a la Ingeniera Lourdes Esquivez Paredes como Asesora Metodológica; y a todas las personas que con las facilidades otorgadas permitieron superar largamente las limitaciones encontradas en la realización del presente trabajo de investigación.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

RESUMEN

En la presente investigación tiene como objetivo proponer una herramienta utilizando las técnicas de minería de datos, donde permita al usuario tener acceso a la información precisa donde se realicen predicciones sobre los alumnos que se matriculen en los próximos años, obteniendo resultados a corto plazo, que permitirá asegurar la confiabilidad de éstos, sirviendo de apoyo a la institución para las decisiones futuras que se puedan tomar.

Dentro de las técnicas predictivas se determinó utilizar los algoritmos de ETS y Redes Neuronales, al realizar el análisis se descartaron algunas técnicas adicionales por no tener los criterios necesarios para su implementación en el modelo a desarrollar.

Además, se presentan los antecedentes de estudio a nivel de base teórica, tomando como fuentes libros, publicaciones, entre otros, los cuales permiten justificar muchos de los conceptos abarcados durante el proceso de investigación.

Se presenta el desarrollo de las metodologías empleadas para la solución del problema planteado; se utilizó la Metodología CRISP DM, como guía para la construcción del modelo de minería de datos basado en series de tiempo logrando realizar las predicciones de deserción escolar en la región de Lambayeque donde solo se tomó como muestra la Ugel de Chiclayo en periodos anuales de manera automatizada dejando de lado el uso de herramientas ofimáticas que retrasan el proceso de los resultados; y el uso de la metodología XP para el desarrollo del sistema como solución a la optimización de los procesos mostrando los resultados

Palabras claves: Minería de Datos, Metodologia Xp, CRISP DM, Ets,Redes Neutronales.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Abstract

In this research it is to propose a tool using data mining techniques, which allow the user to have access to accurate predictions of where students who enroll in the coming years are made, obtaining short-term results, it will ensure the reliability of these, serving as support to the institution for future decisions can be taken.

In predictive techniques, we were determined using algorithms Ets and Neural Networks, since executing some additional analysis techniques were discarded for not having the necessary criteria for its implementation in the model to be developed.

In addition, the background level study theoretical basis, taking as sources books, publications, among others, which can justify many of the concepts covered during the research process are presented.

the development of methodologies for the solution of the problem is presented; CRISP DM Methodology was used as a guide for building the data mining model based on time series managed to obtain predictions dropout in the region of Lambayeque where only took as shown in Chiclayo ugel annual periods so automated aside the use of office tools that slow the process of results; and using the methodology for developing XP system as a solution to the optimization of processes showing the results

Keywords: Data Mining, Methodology Xp, CRISP DM, ETS, Neutronales Networks.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

INTRODUCCIÓN

La deserción estudiantil se ha convertido en un problema social que afecta a muchas Instituciones Educativas en todo el mundo, reducir el número de estudiantes desertores es un tema. Que tienen muy presente cada uno de las Instituciones educativas, donde las mismas planean implementar un plan estratégico para reducir el índice de estudiantes que deciden abandonar sus estudios.

Para contribuir con la solución al problema de la deserción estudiantil se plantea realizar un estudio comparativo de técnicas de minería de datos para predecir la deserción estudiantil en la educación básica regular en la región de Lambayeque.

Donde se seleccionaron las técnicas predictivas para luego compáralas; así mismo también se elaboró una aplicación web usando las técnicas de predicción, para luego evaluar los resultados obtenidos en la investigación.

De acuerdo a (Hand, 2011) , “la Minería de datos es un proceso que reúne un conjunto de herramientas de diversas ciencias, (Estadística, Informática, entre otras)” que persigue extraer conocimiento oculto o información no trivial de grandes volúmenes de datos, con la finalidad de dar soluciones a problemas específicos en las empresas.

CRISM-DM fue la metodología utilizada para la creación del modelo, la misma que es una de las más usadas en la actualidad para la generación de proyectos de Minería de datos, con ella se pretende obtener un modelo de análisis de datos, que con la ayuda de la implementación de algoritmos de Inteligencia Artificial, ya incorporados en la herramienta r-Project, se pueda predecir la probable deserción en las Instituciones educativas y así tomar las medidas preventivas.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Contenido

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Figuras

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Tablas

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Índice de Gráficos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO I

EL PROBLEMA DE INVESTIGACIÓN

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO I: EL PROBLEMA DE INVESTIGACIÓN

1.1. Situación Problemática:

La deserción escolar es un problema que en estos últimos años ha surgido en muchos países. La mayoría de los trabajos que intentan resolver este problema están enfocados en determinar cuáles son los factores que más afectan al rendimiento de los estudiantes, del nivel educativo básico regular, esta realidad es generalizada en los países de Latinoamérica incluyendo el Perú, tal es el caso de (Irizarry & Quintero, 2006), donde la deserción escolar en Puerto Rico es uno de los principales problemas sociales y económicos por una mala planificación familiar donde solo uno de todo el núcleo familiar que la conforma aporta y esto conlleva a que los demás dejen de lado los estudios es por tal motivo que la tasa de deserción se encuentra en un 42%.

Según (Espíndola & León, 2002) señalan que los sistemas educativos de gran parte de América Latina tienen como principal problema una escasa capacidad de retención de los niños y de los adolescentes.

Por otro lado (Elias & Molina, 2005), demuestran que el problema deserción para Paraguay es que a partir de la organización, la comparación y el contraste de las percepciones de educandos y educadores hacen posible que existan inferencias sobre el fenómeno de la deserción escolar. Los puntos emergentes hacen referencia al distanciamiento de la escuela de la realidad del adolescente, los mecanismos de discriminación de género que no son replanteados en la escuela y la violencia contra los propios alumnos que se da dentro de las instituciones escolares.

Por otro lado en Finlandia, no aparece de manera tan problemática ya que cuenta con una tasa del 5% de deserción escolar y una calidad de educación básica mundialmente reconocida. Los factores del por qué en Finlandia la tasa de deserción es baja son diversos pero entre ellos tenemos que cuenta con un sistema educativo que se da de forma gratuita y obligatoria, otro de los factores es la inversión del 6% de su PBI a la enseñanza , llegando así a alcanzar liderar las pruebas Pisa.

Por otro lado en estos últimos años se han realizado investigaciones, congresos, talleres sobre minería de datos tal es el caso (Timarán, Calderón, & Jiménez, 2013) en su investigación, "Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil", donde de los 15.805 registros se seleccionaron únicamente los datos con los atributos más relevantes de los estudiantes de los años 2004 - 2006 , obteniéndose como resultados 6870 registros y 62 atributos correspondientes a la información socioeconómica, académica, disciplinar e institucional. Al finalizar la investigación el autor llego a la conclusión que los factores más relevantes que se determinaron fueron socioeconómicos y académicos asociados a la deserción estudiantil. La minería de datos o data mining se ha ido interrelacionando a lo largo de la vida de cada empresa donde se están interesando en explorar sus bases de datos.

Por otro lado en el Perú, existe un problema fundamental es el identificar y encontrar información útil para poder predecir cuál sería la probabilidad de que un estudiante deserte ya que uno de los inconvenientes que se presenta es la falta de conocimientos de las variables que influyen en la deserción escolar ya que los diversos datos se encuentran por lo general

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

en forma no refinada y para poder analizarlos con fiabilidad es necesario que exista una cierta estructuración y coherencia entre los mismos.

Según (Ministerio de Educación, el 14% de niños y jóvenes entre los 13 y 19 años dejó el colegio o nunca se matriculó, 2014) En el Perú de los 4 millones 300 mil escolares entre los 13 y 19 años de edad. Estas cifras se darían a entender que más del 14 de cada 100 alumnos abandona las aulas en todo el país.

En la mayoría de los casos, las deserciones escolares (de alumnos entre (13 y 19 años) obedecen a problemas económicos (45,1%), desinterés por estudiar (27,2%), problemas familiares (16,6%), y por quehaceres en casa (5,4%).

En Lambayeque, la deserción escolar es un problema preocupante. Solo en el 2013, el total 8.162 alumnos de colegios públicos abandonaron las aulas para trabajar. Esto equivale a un 4.36% del total de escolares lambayecanos que se matricularon en ese mismo año. Por otro lado solo en las ciudades de Chiclayo el ausentismo en las aulas fue alrededor de 4,356 y en la de Ferreñafe fue alrededor de 2,641 escolares.

1.2. Formulación del Problema

La situación actual en el Perú se estimaba que existían alrededor de 3.5 millones de niños y niñas entre 6 y 11 años, edades en las que se debería iniciar y culminar, respectivamente, la educación primaria. (Unicef, s.f.)

De acuerdo con lo expuesto anteriormente se describe en varias investigaciones con respecto al tema donde están tratando de dar solución desde distintos ángulos tal es el caso de (Valero, Salvador, & García, 2003) , Donde en su investigación de Predicción de la deserción escolar,

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

tomando como base de análisis los datos del estudio socioeconómico del EXANI-II, elaborado por el CENEVAL, se usó el algoritmo de árboles de decisiones y el de los k vecinos más cercanos para buscar predecir la deserción escolar en la Universidad Tecnológica de Izucar de Matamoros. Obteniendo como resultado el 70% de aciertos.

(Márquez, Romero, & Ventura, 2012) , en su investigación de Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos, utilizando la técnica de Árbol de decisión, donde buscó predecir la deserción escolar. Los datos que se obtuvieron se recopilaron de tres diferentes fuentes los cuales fueron: una Encuesta, CENEVAL y el departamento escolar de la misma.

Por lo antes expuesto que hasta la fecha cabe destacar que la mayor parte de las investigaciones sobre minería de datos aplicada a los problemas de abandono y fracaso, se han aplicado, sobre todo, en el nivel de educación superior y, en mayor medida en la modalidad de educación a distancia.

En el presente trabajo de investigación se plantea utilizar distintas técnicas de Minería de Datos y compararlas para predecir la deserción estudiantil de la educación regular Lambayecana. Formulando la siguiente pregunta **¿De qué manera las técnicas de minería de datos permiten predecir la deserción estudiantil en la educación básica regular en la región de Lambayeque?**

1.3. Delimitación del Problema:

Este trabajo busca a través de una aplicación web usando las técnicas de minería de datos (redes neuronales y serie de tiempo) para predecir la

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

deserción estudiantil en la educación básica regular (primaria y secundaria) en la región de Lambayeque, tomando como referencia la Ugel Chiclayo.

1.4. Justificación e Importancia de la Investigación

Históricamente la deserción de los alumnos se identifica en el momento que ellos soliciten su baja, sin encontrar claramente las causas a los problemas anteriormente mencionadas; por ello con la presente investigación y a través de la aplicación web que utiliza las diferentes técnicas de minería de datos, para así identificar y calcular el porcentaje de probabilidad de que un alumno pueda desertar y poder aplicar estrategias necesarias para disminuir el índice de deserción.

Por lo tanto, con la investigación se busca realizar un estudio de las técnicas de minería de datos para la predicción de la deserción estudiantil en la educación básica regular y mediante el análisis de las diversas técnicas que existen en el campo de la minería de datos y con los resultados se podrá llevar a cabo una predicción eficiente.

1.5. Limitaciones de la Investigación

Datos incompletos: De los datos obtenidos, se ha observado que algunas encuestas base, no tiene datos en alguno de sus campos, inconsistencia generada por las autoridades de los colegios, que no cumplieron con registrar la información en forma completa.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Incertidumbre: En los diferentes datos obtenidos en formatos pdf, se puede observar que la definición y expresión de los campos requeridos para la investigación no están homogeneizadas, lo que implicara una estandarización en el tipo de dato de la futura base de datos.

Tamaño: La gran cantidad de registros obtenidos, dará lugar a una base de datos de dimensiones considerables en tamaño por el gran número de registros a manejar y la gran complejidad de datos (campos) a definir.

1.6. Objetivo

Objetivo General

Aplicar técnicas de minería de datos para predecir la deserción estudiantil de la educación básica regular Lambayecana.

Objetivos específicos

a) Recopilar y analizar los archivos ofimáticos del nivel básico de la Ugel Chiclayo.

b) Seleccionar las técnicas predictivas de minería de datos.

c) Comparar técnicas de minería de datos a aplicar que mejoren la predicción de la deserción escolar del nivel básico regular.

d) Analizar resultados obtenidos con las diferentes técnicas de minería de datos.

e) Construir un aplicativo web usando las técnicas de predicción.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO II
MARCO TEÓRICO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO II: MARCO TEÓRICO

2.1. Antecedentes de Estudios

Sobre la aplicación de técnicas de minería de datos se han realizado diferentes investigaciones, como tal es el caso de, (Spositto, Etcheverry, Ryckeboer, & Bossero, 2010), La investigación se realizó aplicando el árbol de decisiones (j48) y el algoritmo FT sobre los datos de alumnos del período 2003-2008 para evaluar el rendimiento académico y la deserción de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas sobre los datos de los alumnos del periodo 2003 al 2008. Donde se obtuvieron como resultados con el algoritmo FT un 78,07 % mientras que con el algoritmo j48 72,53% llegando a la conclusión que el algoritmo FT es mejor cuanto al rendimiento escolar es superior al algoritmo j48.

Por otro lado (Valero, Salvador, & García, 2003), en su investigación utilizaron las técnicas de minería de datos para poder predecir la deserción escolar en la Universidad Tecnológica de Izúcar de Matamoros, donde utilizaron los algoritmos tales como: C4.5 y el algoritmo de los k vecinos más cercanos. Obteniendo como resultado que las causas principales que desertan son: La edad, los ingresos familiares, El nivel de inglés. Llegando a la conclusión que con la propuesta planteada podrán determinar los factores de riesgos de manera oportuna.

Por otro lado **(Timarán, Calderón, & Jiménez, 2013),** en su investigación el objetivo fue la detección de patrones de deserción estudiantil partiendo a partir de los datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de la Universidad de Nariño e Institución Universitaria IUCESMAG, donde utilizando técnicas de minería de datos su clasificación estuvo basada en árboles de decisión (j48), donde se seleccionaron los datos socio-económicos, académicos, disciplinares e institucionales de los estudiantes que ingresaron en los años 2004, 2005 y 2006. Obteniendo como resultado que la deserción en la Universidad de Nariño es estrictamente académico. Por lo que llegaron a la conclusión que aplicando las técnicas de clasificación y clustering sobre los datos de los estudiantes se ha obtenido un patrón común de deserción estudiantil, determinado por un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera.

También **(Silvaz Wanumen, 2010).**En su investigación que hizo que lleva por nombre Minería de datos para la predicción de fraudes en tarjetas de crédito uso los algoritmos de árboles de clasificación (j48) y también uso las reglas de asociación (a priori), para la posible detección de fraudes a nivel de tarjetas de crédito.

Donde se compraron los dos algoritmos llegando a la conclusión que las reglas de asociación (a priori), fue menos efectiva que la de clasificación (j48)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.2. Estado del arte

(Timarán, Calderón, & Jiménez, 2013)

Refiere que el objetivo es detectar patrones de deserción estudiantil partiendo de los datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de la Universidad de Nariño e Institución Universitaria IUCESMAG. Utilizando técnicas de minería de datos su clasificación estuvo basada en árboles de decisión (j48), donde se seleccionaron los datos socio-económicos, académicos, disciplinares e institucionales de los estudiantes que ingresaron en los años 2004, 2005 y 2006. Obteniendo como resultado que la deserción en la Universidad de Nariño es estrictamente académico. Por lo que llegaron a la conclusión que aplicando las técnicas de clasificación y clustering sobre los datos de los estudiantes se ha obtenido un patrón común de deserción estudiantil, determinado por un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera.

(Formia, Lanzarini, & Hasperué, 2013)

En el presente trabajo se explica el proceso de identificación de las características más relevantes del problema donde a través , utilizando técnicas de Minería de Datos (DM), puede obtenerse un modelo de la deserción universitaria en la unidad académica mencionada, para lo cual se utilizó el algoritmo de agrupamiento K-medias así se pudo segmentar a los alumnos desertores en grupos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Se eliminaron atributos no generalizables, se redujo la cardinalidad de algunos atributos utilizando categorías más genéricas, se construyeron nuevos atributos mediante funciones de sumarización (summarize), se discretizaron o numerizaron atributos según la necesidad de los algoritmos y se realizaron normalizaciones de rango. Finalmente, se estableció un atributo de estado que diferencia a los alumnos que ya han abandonado (luego1 de un año sin actividad académica) de los que cursan normalmente.

(Barrientos & Ríos, 2013)

En su investigación expone que tenía por objetivo mostrar una metodología para poder predecir la fuga de clientes ó Churn en un ambiente Multiplataforma en la industria de las telecomunicaciones. Además.

El chrun que se calculó en la investigación es aquel referente al servicio debido a que para que sea aplicable a nivel de cliente se debe implementar el KDD como procedimiento relevante en la compañía, además, ésta debe presentar una cultura más orientada a la retención en vez de a la fuerza de venta.

Donde se usaron algunos algoritmos tales como Redes Neuronales, Support Vector Machines y Árboles de Decisión donde se estimó la calidad como el porcentaje de aciertos en la variable predicha.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

(Ortiz Farro, 2015)

En su investigación tuvo como propósito el desarrollo de un sistema Inteligente donde se utilizó técnicas de minería de datos para lo cual se utilizó series de tiempo para poder predecir la producción de arroz. Donde se seleccionaron los datos de año y producción total de los periodos 2001 al 2011.

2.3. Base Teórica Científicas

Se presentan los conocimientos o bases teóricas que serán empleadas a lo largo del desarrollo.

2.3.1. Deserción Escolar

(Bachman, Green, & Wirtanen, 1971), definen que la deserción escolar se ocasiona por aquellos estudiantes que interrumpen su asistencia a la escuela por varias semanas por diferentes razones, exceptuando aquellos por enfermedad.

(Morrow, 1985), define a la deserción cuando un estudiante el cual estuvo inscrito en la escuela, deja la misma por un largo periodo de tiempo y no se inscribió en otro colegio. Donde, no se toman en cuenta,a los estudiantes que estuvieron enfermos o fallecieron.

(Fitzpatrick & Yoels, 1992), se refieren a la deserción, cuando un estudiante deja la escuela sin graduarse, independientemente si regresan o reciben algún certificado equivalente.

(Lavado & Gallegos, 2005), elaboran su propia definición partiendo de las definiciones anteriores, donde llegaron a establecer que la deserción escolar se da siempre y cuando los individuos que habiendo asistido a la

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

escuela el año anterior, en el año actual o corriente no lo están haciendo, exceptuando solo a aquellos que han dejado de asistir por diversos motivos.

Por lo tanto la deserción escolar se define como aquel estudiante que realizo su matrícula o inscripción en un determinado año, y por causas determinadas deja inconclusa su preparación académica.

2.3.2. Minería de Datos

Según (Pérez & Santín, 2008), se refieren inicialmente a la minería de datos como un proceso de descubrimiento de nuevas y significativas relaciones, patrones al examinar grandes volúmenes de datos.

Por otro lado (Carrasco, 2011), expone que la minería de datos es el proceso de extracción de la información de interés partiendo de los datos, donde se entiende que solo el conocimiento es de interés siempre y cuando sea novedoso.

Según (Weiss & Indurkhya, 1998), Define a la minería de datos es la búsqueda de información valiosa en grandes volúmenes de datos. Se trata de un esfuerzo entre los humanos y las computadoras.

2.3.3. KDD: Proceso De Extracción de Conocimiento

Según (Usama & Wierse, 2002), refieren que el KDD es un proceso no trivial para poder identificar patrones válidos, novedosos, potencialmente útiles a partir de los datos.

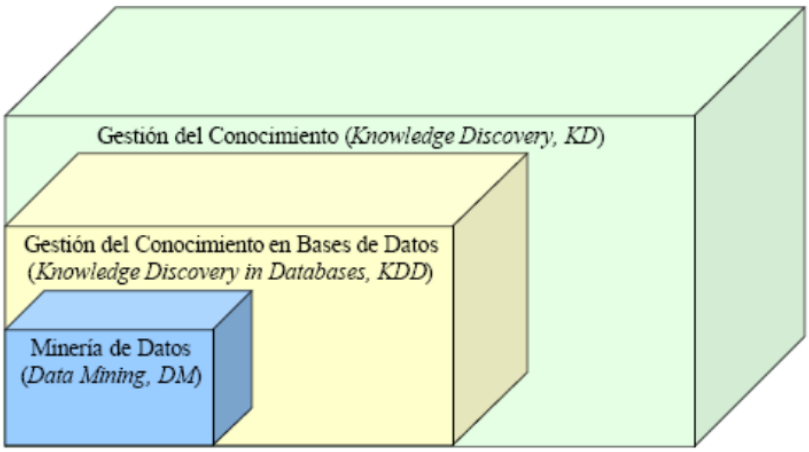
Por otro lado (Guallart Romeu, 2010), contextualizan al KDD como al proceso de búsqueda y extracción de conocimiento partiendo de las bases de datos, mientras que la Minería de Datos es la parte de este

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

proceso en la que se utilizan las técnicas de inteligencia artificial para obtener un modelo.

Hoy en día se puede confundir a la minería de datos con el proceso KDD. Donde la minería de datos forma parte del proceso de KDD como Se puede ver en la Figura. 1 (Guallart Romeu, 2010)

Figura1: Comparación de los conceptos de Minería de Datos, KDD y Knowledge Discovery



Fuente: (Guallart Romeu, 2010)

El KDD forma parte de un área científica más amplia como es el descubrimiento de conocimiento que tiene otras muchas partes dentro de ella diferentes al KDD.

2.3.4. Fases de KDD

Según (Pernía & F., 2001) las fases KDD son:

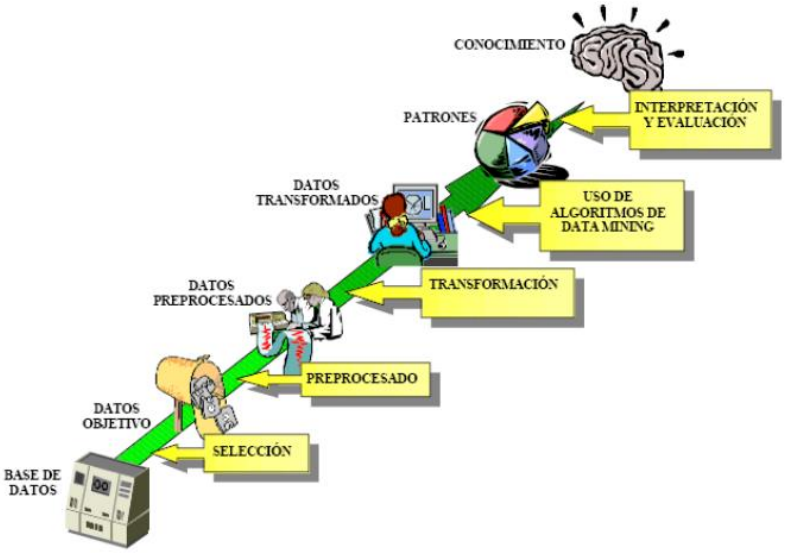
1. Exploración del Dominio.
2. Recolección de los datos
3. Extracción de patrones en los datos
4. Inducir generalizaciones
5. Verificación del conocimiento

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

6. Transformación del conocimiento.

Por otro lado (Brachman & Anand, 1996) define las fases así:

Figura 2: Proceso KDD



Fuente: (Brachman & Anand, 1996)

En (Hernández & Ferri, 2004), en su investigación expone las siguientes fases en el proceso de KDD:

1. Preparar los datos:

a. Especificar las fuentes de información las cuales puedan ser útiles.

b. Elaborar un esquema de almacén de datos (Data Warehouse) para poder unificar toda la información recogida.

c. Implantación del almacén de datos que permita la "navegación" y Visualización previa de sus datos, y asi poder diferenciar que atributos pueden ser interesantes para el estudio.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

d. Selección, limpieza y transformación de los datos que se van a analizar. La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).

2. Minería de Datos:

a. Seleccionar y aplicar el método más apropiado.

b. Evaluación/Interpretación/Visualización.

c. Evaluar, interpretar, transformar y representar los patronos que se extraen.

d. Difundir y uso del nuevo conocimiento que se obtiene.

Según (WebMining Consultores, 2014), las etapas o fases del proceso KDD las divides en 5:

1. **Selección de datos.** En esta fase es determinar cuáles son las fuentes y el tipo de información que se va a utilizar. En esta fase los datos relevantes son extraídos desde la o las fuentes de datos.

2. **Pre procesamiento.** En esta fase se prepara y se limpia los datos que son extraídos desde las distintas fuentes de datos ya que van a ser necesario en las fases posteriores. En esta fase se emplean diversas estrategias para poder manejar datos faltantes, datos inconsistentes o que están fuera de rango, con la finalidad de obtener una estructura adecuada para posteriormente transformarla.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3. **Transformación.** En esta fase consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables partiendo de las existentes con una estructura de datos apropiada. En esta fase se realizan las operaciones de agregación o normalización, donde se consolidan los datos de una forma necesaria para la fase siguiente.

4. **Data Mining**. Es la fase de modelamiento, en donde métodos inteligentes son aplicados con la finalidad de extraer patrones previamente desconocidos, validos, nuevos y potencialmente útiles que están contenidos u ocultos.

5. **Interpretación y Evaluación.** En esta fase es donde se identifican l patrones obtenidos y que son realmente interesantes, y que se basan en algunas medidas y se basándose en algunas medidas y se efectúa la evaluación de los resultados que se obtienen.

2.3.5. Clasificación de las Técnicas de Minería

En tanto la clasificación de la minería de datos entre autores se difiere:

Según, (Joshi, 1997)**,** los componentes de la minería de datos son los siguientes:

1. **Clustering:** Donde se analizan los datos y se generan conjuntos de reglas que agrupan y clasifican los datos futuros.

2. **Reglas de asociación:** Son aquellas reglas o condiciones que presentan un grupo de objetos de una base de datos un ejemplo de regla de asociación o condición sería: “Un 30% de las transacciones que contienen toallitas de bebé, también contienen pañales; 2% de las transacciones contienen toallitas de bebé”. En el ejemplo antes

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

mencionado el 30% es el nivel de confianza de la regla y 2% es la cantidad de casos que respaldan la regla.

3. **Análisis de secuencias:** Trata de descubrir patrones que suceden en una Secuencia determinada. Trabaja sobre datos que se presentan en Distintas transacciones. “Muchos usuarios que han comprado X luego Han comprado Y”.

4. **Reconocimiento de patrones:** Analiza la asociación de una señal de Información de entrada con aquella o aquellas con las que guarda mayor similitud, de entre las catalogadas por el sistema. Se usan para identificar causas de problemas o incidencias y buscar posibles soluciones, siempre y cuando se adecua a la base de información necesaria en donde buscar.

5. **Predicción:** Se busca determinar el comportamiento futuro de una variable o un conjunto de variables a partir de la evolución pasada y presente de las mismas o de otras de las que dependen. Las técnicas asociadas a estas herramientas tienen ya un elevado grado de madurez.

6. **Simulación:** Comparan la situación actual de una variable y su posible evolución futura.

7. **Optimización:** Resuelve el problema de la minimización o maximización de una función que depende de una serie de variables.

8. **Clasificación:** Permiten asignar a un elemento la pertenencia a un determinado grupo o clase. Se establece un perfil característico de cada clase y su expresión en términos de un algoritmo o reglas, en función de distintas variables. Se establece también el grado de discriminación o

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

influencia de estas últimas. Con ello es posible clasificar un nuevo elemento una vez conocidos los valores de las variables presentes en él.

Mientras que para (Cabena, 1998) , compone a la minería de datos en cuatro grandes operaciones soportadas por algunas técnicas comúnmente usadas

1. **Modelización predictiva**: Que usa las técnicas de:
    a) Clasificación
    b) Predicción de valores
2. **Segmentación de bases de datos:** Que usa técnicas de:
    a) Clustering poblacional
    b) Clustering por redes neuronales
3. **Análisis de relaciones:** Que utiliza las técnicas de:
    a) Descubrimiento de asociaciones
    b) Descubrimiento de secuencias de patrones
    c) Descubrimiento de secuencias temporales similares
4. **Detección de desviaciones:**
    a) Técnicas estadísticas
    b) Técnicas de visualización

Según (Guallart Romeu, 2010) se puede clasificar las técnicas de aprendizaje de la siguiente manera:

1. **Métodos inductivos**: Son aquellos que partiendo de los datos iniciales y del conocimiento generado son capaces de construir modelos que a partir de los datos generen los resultados.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2. **Técnicas predictivas:**

Interpolación: Es la generación de una función continua sobre varias dimensiones.

Predicción secuencial: Es cuando las observaciones están ordenadas en forma secuencial y se puede predecir el siguiente valor de la secuencia.

Aprendizaje supervisado: En éstas técnicas cada observación, compuesta por muchos valores de atributos, donde se interpone un valor de la clase a la que corresponde. Se genera un clasificador a partir de clases que se proporcionan. Es un caso particular de interpolación en el que la función genera un valor discreto en lugar de continuo

3. **Técnicas descriptivas:**

Aprendizaje no supervisado: Es el conjunto de observaciones las cuales no tienen algunas clases asociadas. Tiene como objetivo la detección regularidades en datos de cualquier tipo: agrupaciones de datos parecidos o próximos, contornos de delimitación de grupos, asociaciones o valores anómalos.

Métodos abductivos: Se pretende, partiendo de los valores generados y de las reglas, obtener los datos de origen. El objetivo es la explicación de evidencia con respecto a los sucesos que se han producido, tal cual haría un investigador privado, que a partir de las consecuencias de los hechos y de ciertas reglas

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

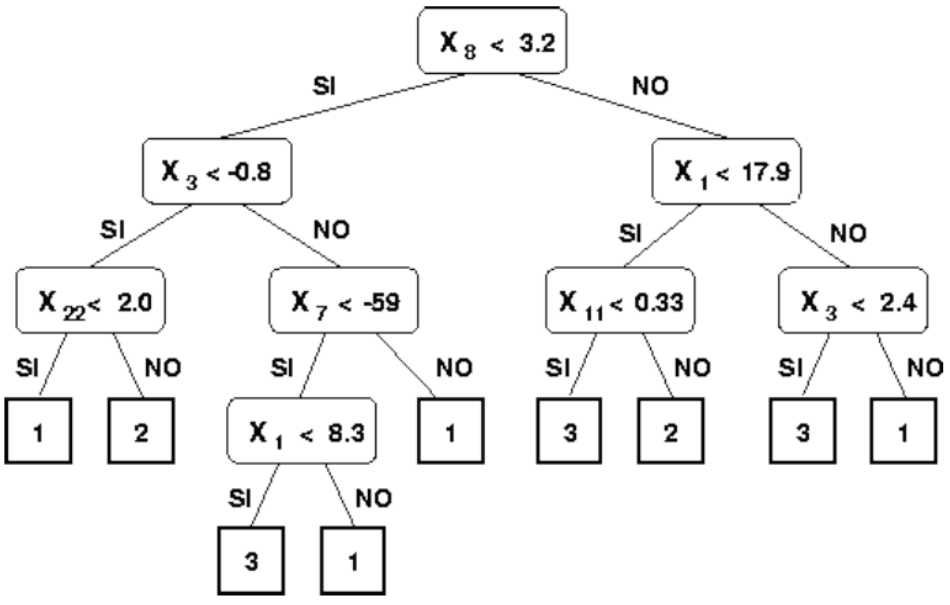
2.3.6. Técnicas de minería de Datos

2.3.6.1. Arboles de decisiones

Los árboles de decisión son una de las formas más populares de Minería de Datos porque tienen una representación sencilla de problemas con un número finito (y a ser posible reducido) de clases. Además son modelos comprensibles y proposicionales (Hernández & Ferri, 2004)**.**

Un claro ejemplo de un árbol de decisión en (Guallart Romeu, 2010) .Donde partir del valor de la variable X8, si el valor es menor de 3.2 se continuará la toma de decisiones por la rama izquierda y si es mayor o igual se continuará por la rama de la derecha. A partir de aquí cada rama tiene una variable separadora con un valor de separación, y así sucesivamente formando un árbol.

Figura 3: Árbol de decisión



Fuente: (Guallart Romeu, 2010)

Donde también **(Mazo & Bedoya, 2010)**, puntualizan que un árbol de decisión es una estructura en la cual cada nodo interno significa una prueba sobre uno o varios atributos, donde cada rama representa una salida de la prueba y los nodos hojas representan clases.

2.3.6.2. C4.5

Según (Quinlan, 1993)**,** y su versión comercial C5.0 Es una extensión de ID3, el cual permite que se trabaje con valores continuos para los atributos, donde se separan los resultados en dos ramas: una para aquellos Ai=N y la otra para Ai>C4.5, donde es capaz de trabajar con ejemplos que contienen valores desconocidos y es tolerante a datos con ruido.

2.3.6.3. Métodos Bayesianos

Una de las características primordiales de los métodos bayesianos es el uso de distribuciones de probabilidad para cuantificar incertidumbre de los datos que se desea modelar. Estos métodos proporcionan una metodología práctica para la inferencia y predicción y, en última instancia, para tomar decisiones que involucran cantidades inciertas (Hernández & Ferri, 2004)

(Hernández & Ferri, 2004) **,** dice que “es una de las que más se han utilizado en problemas de inteligencia artificial, con ello en el aprendizaje automático y minería de datos, ya que es un método práctico para realizar inferencias a partir de los datos, la misma que se basa en estimar la probabilidad de pertenecía (a una clase o grupo) mediante la estimación de las probabilidades, utilizando para ello el teorema de Bayes”.

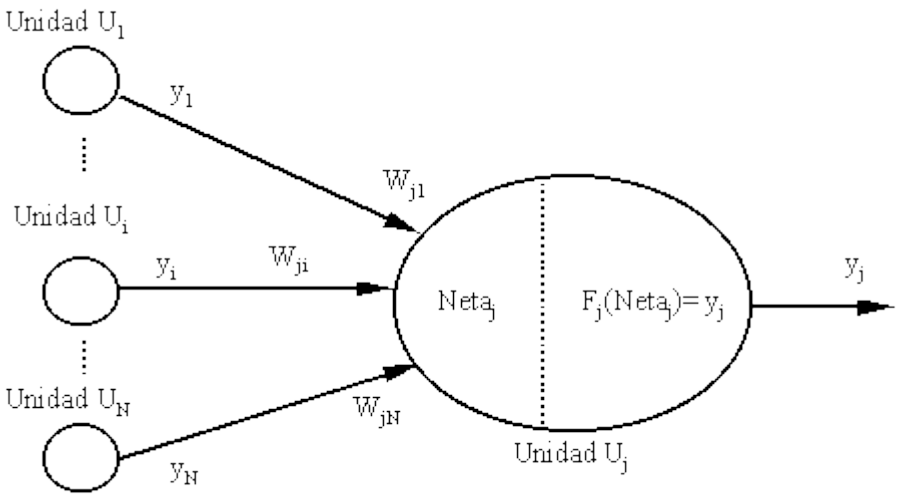
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.3.6.4. Redes neuronales artificiales

Según (Hernández & Ferri, 2004) señala que las redes neuronales posee dos tipos de aprendizaje uno es el supervisado, en el mismo que se le proporciona un conjunto de datos de entrada y la respuesta correcta es útil en tareas de regresión y clasificación. Y el aprendizaje no supervisado solo se le da a la red un conjunto de datos de entrada y la red debe auto-enseñarse para proporcionar una respuesta, este aprendizaje es útil para las tareas de agrupamiento.

Donde las redes neuronales han sido utilizadas en diversas áreas de estudio tal es el caso en la predicción de mercados financieros, control de robots, etc. (Guallart Romeu, 2010)**.**

Figura 4: Esquema de una Neurona Artificial con sus principales Elementos



Fuente: (Lopez Alfonso, 2015)

En redes neuronales hay dos tipos principales de aprendizaje en RNA:

a) Aprendizaje supervisado: Estos algoritmos precisan que cada vector de entrada se empareje con su correspondiente vector de salida. Mientras que el entrenamiento se basa en la de mostrar un vector de entrada a la red, donde se calcula la salida de la red y después se compara con la salida deseada y por otro lado el error o diferencia resultante se emplea para realimentar la red y modificar los pesos de acuerdo con un algoritmo que tiende a minimizar el error. (Olabe Basogain, 2008)

b) Aprendizaje no supervisado: Son aquellos sistemas donde al aprendizaje solo se le da un determinado conjunto de datos de entrada y la red debe auto-enseñarse y así proporcionar una respuesta, donde este aprendizaje es de gran utilidad para tareas de agrupamiento. (Olabe Basogain, 2008).

2.3.6.5. K –means

Este algoritmo es uno de los más utilizados con lo que respecta al agrupamiento de datos, es el K-Medias o también conocido como K-Means por ser uno de los más veloces y eficaces. El algoritmo trabaja con un método de agrupamiento por vecindad, en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar sin etiquetar.

El propósito de K-Means es ubicar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares. (Moody & Darken, 1989)

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Todo ejemplo nuevo, una vez que los prototipos han sido correctamente situados, es comparado con estos y asociado a aquel que sea el más próximo, en los términos de una distancia previamente elegida.

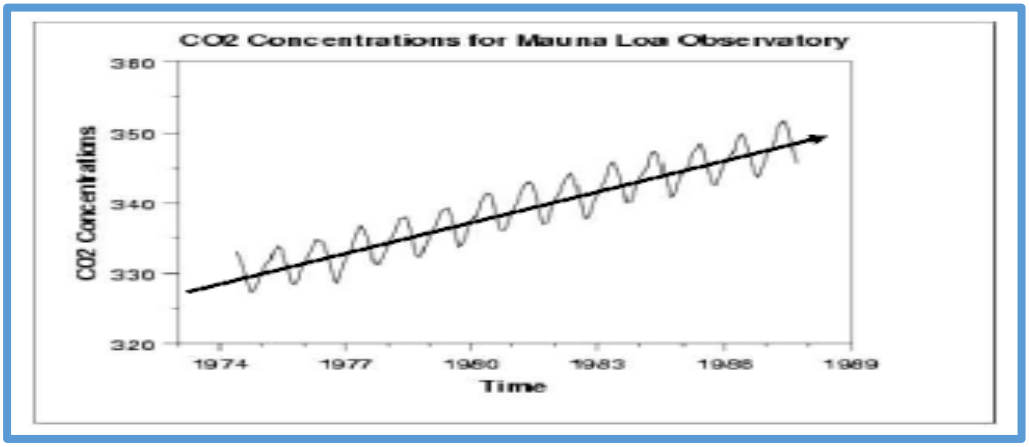
Normalmente, se utiliza la distancia euclidiana. El objetivo que se busca mediante el algoritmo K-Means es minimizar la varianza total intragrupo o la función de error cuadrático, para que el algoritmo pueda generar los mejores resultados.

2.3.6.6 Series de tiempo

Es aquel conocimiento que se obtiene atrevés de la recopilación de datos, la observación o el registro de intervalos de tiempos regulares, donde que a partir de ese conocimiento y con el supuesto de que no se producirán cambios, y así poder realizar predicciones. Algunas definiciones que se usan con esta técnica son:

A) **Tendencia:** Es aquel componente a largo plazo la cual representa la disminución o crecimiento en un amplio periodo de tiempo.

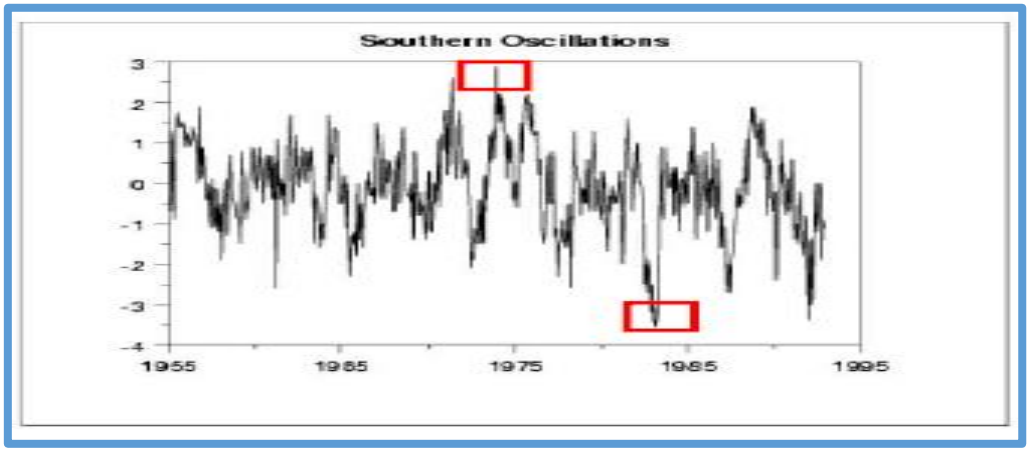
Figura 5: Gráfico de Tendencia de un conjunto de datos de los años 1974-1989.



Fuente: **(Cruz Arrela, 2010)**

B) **Estacionalidad**: Es aquel elemento en el cual se presenta en series de frecuencia inferior a la anual, y se presume oscilaciones a un corto plazo regular, inferior al año y amplitud regular.

Figura 7: Gráfica de valores en el tiempo, donde se observa la estacionalidad



Fuente: **(Cruz Arrela, 2010)**

C) ETS (Exponential smoothing state)

(Hyndman R. J., 2014), Los Métodos de suavización exponencial han existido desde la década de 1950, y son los métodos de pronóstico más populares utilizados en los negocios y la industria. Recientemente, suavizado exponencial ha revolucionado con la introducción de un marco de modelización completa incorporando innovaciones modelos de estado espacio, cálculo de probabilidades, los intervalos de predicción y los procedimientos para la selección del modelo.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ETS (M, N, N) Suavización exponencial simple con errores multiplicativos: Según (Hyndman,2014) se puede especificar modelos con errores multiplicativos escribiendo los errores aleatorios de un solo paso como errores relativos:

$$\varepsilon_t = \frac{y_t - \hat{y}_{t|t-1}}{\hat{y}_{t|t-1}}$$

where $\varepsilon_t \sim \text{NID}(0, \sigma^2)$. Substituting $\hat{y}_{t|t-1} = \ell_{t-1}$ gives $y_t = \ell_{t-1} + \ell_{t-1}\varepsilon_t$ and $e_t = y_t - \hat{y}_{t|t-1} = \ell_{t-1}\varepsilon_t$.

Entonces se puede escribir la forma multiplicativa del modelo de espacio de estados como se muestra:

$$\begin{aligned} y_t &= \ell_{t-1}(1 + \varepsilon_t) \\ \ell_t &= \ell_{t-1}(1 + \alpha\varepsilon_t). \end{aligned}$$

D) Holwinters:

Es una variante, donde es conocida como alisado exponencial lineas con doble parámetro,donde consigue la eliminacion del sesgo de la predicción de una serie de tendencia , a través de la inclusión en la media móvil de un componente de tendencia.

Por otro lado comprando con diversas técnicas, tal como ARIMA, donde el tiempo necesario para el calculo en la predicción es considerablemente rápido.

De hecho, Holt-Winters es utilizado por diversas compañías para el pronóstico de la demanda a corto plazo siempre y cuando los datos de venta contengan tendencia y patrones estacionales de un modo subyacente.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

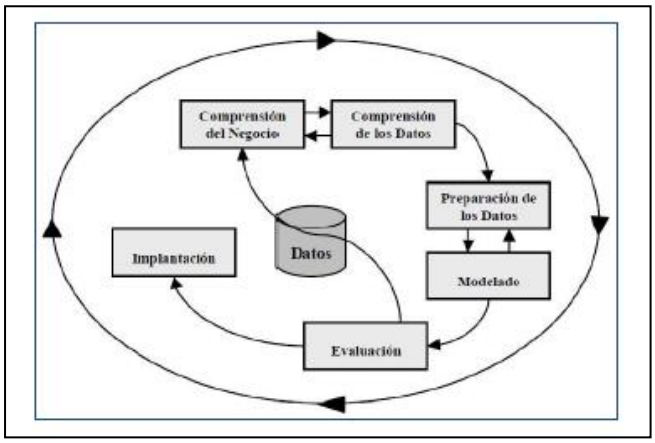
2.3.7. Metodologías para la aplicación de minería de datos

a) CRISP – DM (Cross Industry Standard Process for Data Mining)

CRISP-DM organiza el desarrollo de un proyecto de Data Mining en una serie de fases o etapas, con tareas generales y específicas que permitan cumplir con los objetivos del proyecto. Estas fases funcionan de manera Cíclica e iterativa, pudiendo regresar desde alguna fase a otra anterior.

Se basa en función a un modelo jerárquico de procesos, donde se establece un ciclo de vida de los proyectos de explotación de información

Figura 7: Fases del modelo CRISP - DM



Fuente: "CRISP – DM 1.0: Step by Step Data Mining guide".

Según (Orallo Hérnandez, 2015) las fases de la metodología crisp son las siguientes:

a. Comprensión del negocio: Es donde se infiere tanto como los objetivos y requerimientos del proyecto desde una perspectiva de negocio.

b. Comprensión de los datos: Se selecciona y adapta los datos, para poder identificar los problemas de calidad de datos y así obtener datos potenciales para poder analizar.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

c. Preparación de los datos: Transformación de los datos. Se seleccionan los datos a utilizar y éstos pasan a una fase de limpieza, estructuración, integración y formateo.

d. Modelamiento y evaluación: Selección y aplicación de Data Mining e Interpretación y evaluación. Se selecciona la técnica a utilizar, construyendo el modelo, para luego ser sometido a diferentes pruebas y evaluaciones.

e. Despliegue del proyecto: Es donde se explota todo el potencial de los modelos y así intégralos en los procesos de toma de decisión de organización, y así difundir el conocimiento extraido, etc.

b) SEMMA

La metodología semma se caracteriza principalmente por la que toma su nombre de las etapas que esta metodología define para procesos de explotación de información, estas etapas son: **muestreo** (sample), **exploración** (explore), **modificación** (modify), **modelado** (model) y **valoración** (assess).

La metodología semma fue desarrollada por la empresa SAS Institute Inc.,una de las mayores organizaciones relacionadas con el desarrollo con el software de inteligencia de negocios.SEMMA esta desarrollada para aplicarla sobre la herramienta de minería de datos “SAS Entreprise Miner”.

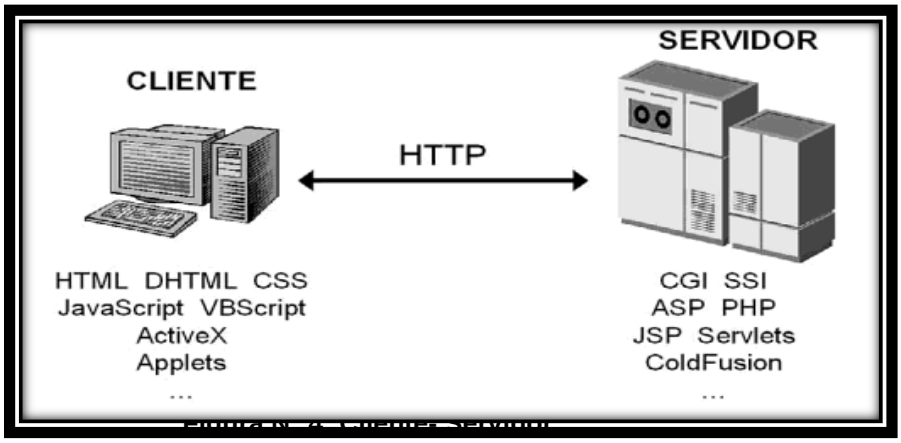
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2.3.8. Aplicación web

Según (Mora, 2002), afirma que “una aplicación web es un tipo especial de aplicación cliente/servidor, donde el cliente, el servidor y el protocolo mediante el que se comunican, están estandarizados y no han de ser creados por el programador de aplicaciones” (p.48).

Según (Mora, 2002), afirma que “suelen distinguirse en tres niveles: el nivel superior que interacciona con el usuario el nivel inferior que proporciona los datos y el nivel intermedio que procesa los datos.”(p.47).

Figura 8: Cliente-Servidor



Fuente: “Programación de aplicaciones web: Historia, principios básicos y clientes web.”

2.3.9. Herramientas de Minería de datos

Según (oocities, s.f.) , para la aplicación de técnicas de minería de datos se clasificaría en dos Librerías y herramientas específicas:

Donde las librerías de Minería de datos son un conjunto de métodos donde se implementan funcionalidades y utilidades básicas como el acceso a datos, modelos de redes neuronales, métodos bayesianos, exportación de resultados Las librerías se encargan principalmente de facilitar el desarrollo

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

de las tareas de Minería de Datos que son más complejas, como el diseño de experimentos. El problema de las librerías, es que es precisa la comprensión de conocimientos de programación.

Algunas de las Librerías más importantes son:

1. **Xelopes (Extended Library For Prudys Embedded Solution):** Es una librería bajo la licencia pública GNU para el desarrollo de aplicaciones de Minería de Datos. Esta librería está implementada para que sea eficiente para la mayoría de los algoritmos de aprendizaje, por eso, es importante destacar que el usuario puede desarrollar aplicaciones particulares de Minería de Datos. Sus principales características son:
    1. Acceso a datos
    2. Modelos de redes neuronales
    3. Métodos de agrupamiento
    4. Métodos de reglas de asociación
    5. Árboles lineales
    6. Árboles no lineales
2. **Mlc++ (Machine Learning Library In C++):** Es un conjunto de librerías que fueron desarrolladas por la Universidad de Standford. La mayoría de las versiones son bajo dominio de investigación, a excepción de la versión 1.3.x, que se distribuye bajo licencia de dominio público. Las principales características son:
    1. Acceso a datos.
    2. Transformaciones de datos
    3. Métodos de aprendizaje mediante objetos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3. **Suites:** Posee las mismas capacidades que el procesamiento de datos, los modelos de análisis, el diseño de experimentos o el soporte gráfico para la visualización de resultados. En este caso, Suites destaca porque existe una interfaz que facilita la interacción entre el usuario y la herramienta.
4. **R-Project:** Es un entorno de trabajo basado en los entornos de programación S y S-PLUS desarrollados a principios de los años noventa del pasado siglo por Bill Venables y David M. Como señalan Venables et al. (2011), es un entono integrado de facilidades informáticas para la manipulación de datos, el cálculo y la generación de gráficos. R-Project pretende convertirse en un sistema internamente coherente que se caracterizaría por un desarrollo basado en la contribución relativamente altruista de la comunidad científica. (López Puga, 2010)
5. **Spss Clementine:** Es uno de los sistemas de Minería de Datos más conocidos. Posee una herramienta visual desarrollada por ISL que tiene una arquitectura cliente / servidor. Este sistema se caracteriza por:
    1. Acceso a datos.
    2. Procesamiento de Datos.
    3. Técnicas de Aprendizaje.
    4. Técnicas de evaluación de modelos.
    5. Visualización de resultados.
    6. Exportación.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

6. **Weka (Waikato Environment For Knowledge Analysis):** Es una herramienta visual de libre distribución desarrollada por los investigadores de la Universidad de Waikato en Nueva Zelanda. Sus principales características son:
    1. Acceso a los datos desde un archivo en formato ARFF.
    2. Pre procesado de datos.
    3. Modelos de Aprendizaje.
    4. Visualización del entorno.
7. **Kepler:** Sistema desarrollador y transformado en una herramienta comercial distribuida por Dialogis. Posee múltiples modelos de análisis. Sus principales herramientas de aprendizaje son:
    1. Árboles de decisión.
    2. Redes neuronales.
    3. Regresión no lineal.
    4. Aplicaciones estadísticas.
8. **Odms (Oracle Data Mining Suite):** Está diseñado sobre una arquitectura cliente servidor; ofrece una gran versatilidad en cuanto al acceso a grandes volúmenes de información. Se caracteriza principalmente por:
    1. Acceso a datos en diversos formatos: almacenes de datos, bases de datos relacionales como SQL, Oracle,etc.
    2. Pre procesado de datos: muestreo de datos, patrones de datos.
    3. Modelos de aprendizaje: redes neuronales, regresión lineal.
    4. Herramientas de visualización.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

9. **Yale:** herramienta de aprendizaje automático implementado en Java por la Universidad de Dormund. El sistema incluye operaciones para:
    1. Importación y pre-procesamiento de datos
    2. Aprendizaje automático
    3. Validación de modelos

2.4. Definición De Términos Básicos

2.4.1. Método

Modo ordenado y sistemático de proceder para lograr un fin / conjunto de reglas (Getoor & Ben, 2007)

2.4.2. Metodología

Conjunto de métodos que se siguen en una disciplina científica / ciencia del método y de la sistematización científica. (Grudnitsky, 1992)

2.4.3. Predicción

Es la acción de aquello que supuestamente va ocurrir. Donde se puede predecir partiendo de conocimientos científicos, revelaciones o de algún tipo de indicios. (Española)

2.4.4. Deserción Escolar

(Bachman, Green, & Wirtanen, 1971), Refieren que la deserción escolar se originan siempre y cuando aquellos estudiantes irrumpen su asistencia al colegio por varias semanas.

2.4.5. Minería De Datos

(Sinnexus, s.f.), exponen que minería de datos es un conjunto de técnicas y tecnologías donde permitirían explorar grandes bases de datos, de manera

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

automática, donde tiene como objetivo el encontrar patrones repetitivos, para así poder explicar el comportamiento de los datos en un contexto determinado.

2.4.6. Técnicas De Predicción

(Universidad de Barcelona, s.f.), se refiere que es lograr la obtención de estimaciones de una serie temporal partiendo de su información histórica inicial hasta la actualidad.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO III

MARCO METODOLOGICO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPÍTULO III: MARCO METODOLÓGICO

3.1. Tipo y diseño de la investigación

La presente investigación es de tipo Tecnológica y diseño Experimental.

Es tecnológica porque a través del uso científico se buscan aplicaciones prácticas (investigación aplicada) para el uso de un producto o también el mejoramiento del mismo.

Es cuasi experimental, ya que se buscara dar explicación como la variable independiente influirá en la variable dependiente.

3.2. Población y muestra

Población

El elemento de estudio determinado como población es el Elemento de Registro en los períodos 2006 – 2015 en la Región Lambayeque donde está conformado por tres ugeles: Chiclayo,Lambayeque y ferreñafe.

Muestra

Ugel Chiclayo

3.3. Hipótesis

La deserción estudiantil de la educación básica regular se puede predecir usando técnicas de minería de datos.

3.4. Operacionalización

Variable independiente

Técnicas predictivas de minería de datos.

Variable dependiente

Predicción de la deserción estudiantil.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 1: Operacionalización de variables

VARIABLES	DIMENSIONES	INDICADORES	ÍTEMS O RESPUESTAS
DEPENDIENTE *Predicción de la deserción estudiantil*	Predicción	Confiabilidad de la predicción.	CP = # Pruebas sin error / Total Población Registrado.
		Tiempo para generar estimación.	TS=Tiempo en segundos
INDEPENDIENTE *Técnicas predictivas de minería de datos.*	Técnica	Tiempo de Procesamiento del Modelo.	TPM = Técnica Modelo Anterior -Técnica del Modelo Propuesto.

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.5. Métodos, técnicas e instrumentos de recolección de datos

3.5.1. Métodos de la Investigación:

En la presente investigación el método de investigación que se utiliza son: la observación, análisis, síntesis y experimental.

a) Observación, Son los análisis que puedo realizar yo mismo, asesores y jurado calificador de la presente investigación.

b) Síntesis, Porque una vez que se analizado el problema planteado y los métodos de visión artificial a implementar, se plantea a desarrollar una solución bajos lo métodos que se han seleccionados.

c) Análisis, Porque se tiene que descomponer el objeto de estudio en sus partes para conocer sus riesgos y propiedades.

d) Experimental, Puesto que se ejecuta a partir de una situación real de un problema, abordándose en la implementación de métodos de visión artificial en la cual fundamento la elaboración y verificación de la hipótesis.

3.5.2. Técnicas de la Investigación

Las técnicas de investigación que se utiliza en el estudio son el análisis y observación.

a. Análisis documental, Consiste en extraer la información de los diferentes, libros, papers, artículos, los cuales presentan una serie de teorías, técnicas, métodos que dan solución a determinados problemas. Todo servirá para limitar la investigación y caracterizar el modelo a estudiar, para analizar resultados obtenidos con las técnicas aplicadas.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

b. Observación: Es el registro visual de lo que ocurre en una situación real, donde se clasifican los acontecimientos con algún esquema y dependiendo el problema que se estudia. En esta técnica es debido está atento para determinar de una forma adecuada todos los resultados confiables de las predicciones.

3.5.3. Instrumento de la Investigación

Cuadro resumen de predicciones: Se realizará una ficha que servirá para recopilar los resultados predictivos que se obtendrán a partir de las técnicas de minería de datos aplicadas en la investigación.

3.6. Procedimiento para la recolección de datos

Para el desarrollo de la presente investigación, está basado en la utilización de técnicas de minería de datos donde se compone de los siguientes pasos:

1. Recopilación De Datos. En esta fase es donde se recolecta toda la información disponible. Para lo cual en primer lugar se debe de seleccionar el conjunto de factores que puedan afectar esto primero el conjunto de los factores que puedan afectar y después se deberán recoger a partir de las diferentes fuentes de datos disponibles. Finalmente toda la información se deberá integrar en un solo y único conjunto de datos

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

2. **Pre-Procesado**. En esta fase es donde se prepara los datos para poder así posteriormente, aplicar las diversas técnicas de minería de datos se deberá preparar los datos para poder aplicar posteriormente, las diversas técnicas de minería de datos. Para ello se deberán realizar las tareas de pre-procesado tales como: limpieza de datos, transformación de variables y particionado de datos, donde también se aplican otras técnicas como la selección de atributo y el re-balanceado de datos para así poder intentar dar solución a los problemas de alta dimensionalidad y desbalanceo que se presentan en este tipo de conjunto de datos.

3. **Minería De Datos**. En esta fase se aplicaran los diversos algoritmos de minería de datos para poder predecir la deserción escolar como si fuera un problema de clasificación. Donde finalmente, los diversos algoritmos empleados deberán ser evaluados y comparados para luego establecer cuál de ellos obtiene el mejor resultado.

4. **Interpretación De Los Resultados**. En esta última fase, es donde se analizan los modelos que obtuvieron unos resultados óptimos para predecir la deserción.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.7. Análisis Estadístico e interpretación de los datos

El análisis Estadístico de datos se basa en lo siguiente:

3.7.1 En el uso de tablas, para la evaluación de las técnicas predictivas en minería de datos.

3.7.2 En el uso de gráficos estadísticos, para la evaluación de las técnicas predictivas en minería de datos

3.7.3 Tiempo de Procesamiento del Modelo, denotada por “TPM”, es el resultado de la Técnica del Modelo Anterior en comparación a la Técnica del Modelo Propuesto:

$$tpm = \frac{tma}{tmp}$$

Dónde:

tpm = Tiempo de Procesamiento del Modelo.

tma = Técnica del Modelo Anterior.

tmp = Técnica del Modelo Propuesto

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

3.8. Criterios de rigor científico

Criterios	Características éticas de los criterios
Confidencialidad	Asegurar la protección de identidad de sus fuentes, como también de las personas que participan como informantes de la investigación.
Manejo de Riesgos	La investigación requiere de una eficiencia y no de un beneficio personal para realizar una investigación consistente.
Observación Participante.	La participación los tesistas requiere una responsabilidad ética por los efectos y consecuencias que pueden surgir durante la investigación.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO IV
ANALISIS E INTERPRETACION DE LOS RESULTADOS

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO IV: ANALISIS E INTERPRETACION DE LOS RESULTADOS

4.1. Resultados

A. Confiabilidad de la Predicción

En este indicador mide el grado de confianza de cada algoritmo seleccionado, dado los objetivos de la presente investigación donde se debe evaluar las técnicas, en este caso son: Redes Neuronales y ETS.

$$PCP = 100 - (\frac{\sum \frac{MR - MP}{MR}}{N} * 100)$$

PCP: Porcentaje de confiabilidad de predicción.

MP: Monto pronosticado

MR: Monto real

N: Número de observaciones

Tabla 2: Generación de Pronósticos Primaria

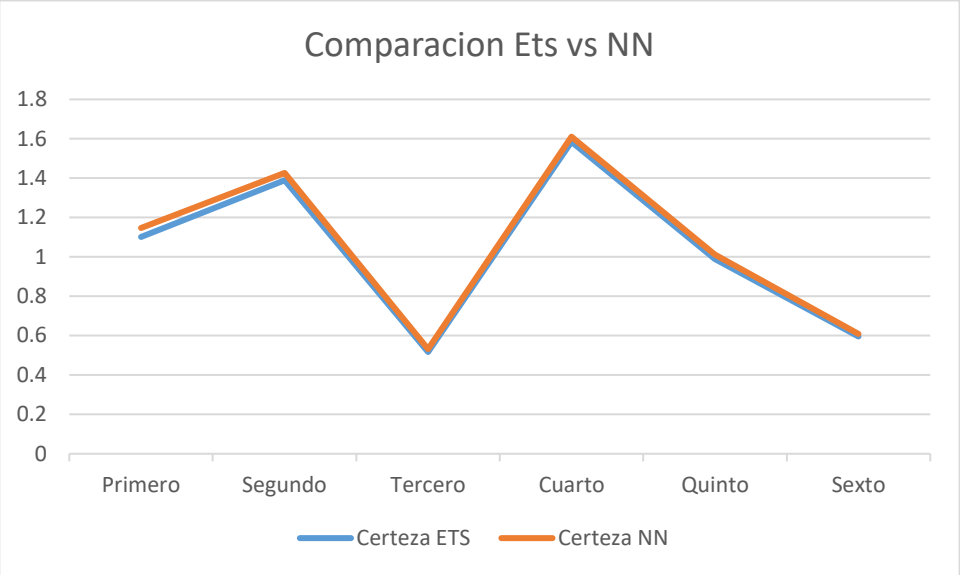
Colegio	Año	Nivel	Grado	Consolidado Matricula	ETS	Red Neuronal
276188	2015	Primaria	Primero	109	120	125
277098	2015	Primaria	Segundo	108	150	154
278658	2015	Primaria	Tercero	269	139	143
278601	2015	Primaria	Cuarto	154	244	248
278516	2015	Primaria	Quinto	175	173	177
278658	2015	Primaria	Sexto	309	184	188

Fuente: Elaboración Propia

En la tabla 2 se muestra los resultados obtenidos del pronóstico generado para el nivel primario utilizando los algoritmos: Red neuronal y ETS en comparación al consolidado de Matricula del año 2015.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Grafico 1: Pronósticos de Matriculas: ETS y Red Neuronal-Secundaria



Fuente: Elaboración propia

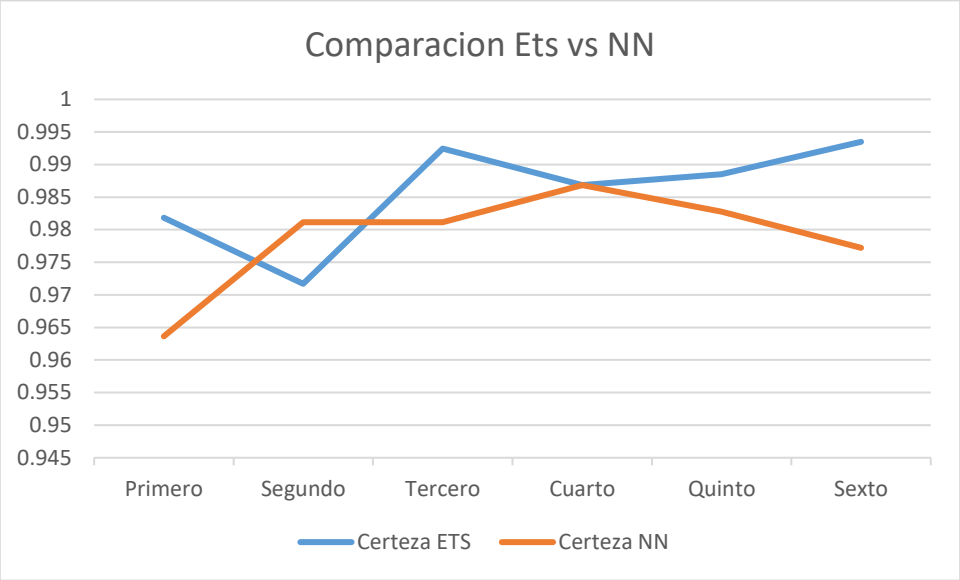
En el gráfico N° 1 podemos observar gráficamente la comparación entre las dos técnicas NN y ETS.

Tabla 3: Generación de Pronósticos Secundaria

Colegio	Año	Nivel	Grado	Consolidado Matricula	ETS	Red Neuronal
276188	2015	Secundaria	Primero	167	108	106
277098	2015	Secundaria	Segundo	242	103	104
278658	2015	Secundaria	Tercero	194	263	260
276032	2015	Secundaria	Cuarto	193	150	150
674187	2015	Secundaria	Quinto	158	172	171

En la tabla 03 se muestra los resultados obtenidos del pronóstico generado para el nivel Secundario utilizando los algoritmos: Red neuronal y ETS en comparación al consolidado de Matricula del año 2015.

Grafico 2: Pronósticos de Matriculas: ETS y Red Neuronal-Primaria



Fuente: Elaboración Propia

En el gráfico N° 2 podemos observar gráficamente la comparación entre las dos técnicas NN y ETS.

Tabla 4: Resultados Obtenidos del nivel secundario aplicando formula

Colegio	Año	Nivel	ETS	Red Neuronal
276188	2015	Secundaria	1.81818182	3.63636364
277098	2015	Secundaria	2.83018868	1.88679245
278658	2015	Secundaria	0.75471698	1.88679245
278601	2015	Secundaria	1.31578947	1.31578947
278516	2015	Secundaria	1.14942529	1.72413793
278658	2015	Secundaria	0.6514658	2.28013029
		Total	8.51976804	12.7300062

Fuente: Elaboración propia

Grado de confianza=100-Total

Red Neuronal	**91.48023196**	**ETS**	**87.2699938**

En la tabla 4, observamos que de acuerdo a la fórmula aplicada el porcentaje de confiabilidad del modelo con respecto a los pronósticos para el nivel secundario arrojados en los años determinados en la muestra

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

obteniendo un grado de confianza en ETS se obtuvo el 87.27%,contra la red Neuronal que obtuvo un 91.48%. Por lo tanto el nivel de confianza más elevado corresponde a la red neuronal con respecto a ETS.

Tabla 5: Resultados Obtenidos del nivel Primario aplicando formula

Colegio	Año	Nivel	ETS	Red Neuronal
276188	2015	Primaria	0.92592593	1.83486239
277098	2015	Primaria	0.93457944	1.85185185
278658	2015	Primaria	0.37313433	0.74349442
276032	2015	Primaria	0.52083333	0.51813472
674187	2015	Primaria	0.63694268	1.26582278
		Total	8.51976804	6.21416616

Fuente: Elaboración propia

Grado de confianza=100-Total

Red Neuronal	96.6085843	ETS	93.78583384

En la tabla N° 05, observamos que de acuerdo a la fórmula aplicada el porcentaje de confiabilidad del modelo con respecto a los pronósticos para el nivel primario arrojados en los años determinados en la muestra obteniendo un grado de confianza en Red Neuronal se obtuvo el 96.60%,contra ETS que obtuvo un 93.76%. Por lo tanto el nivel de confianza más elevado corresponde a la red neuronal con respecto a Red Neuronal.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

B. Tiempo de Procesamiento del Modelo.

Este indicador mide el tiempo que le toma a cada técnica calcular u obtener la estimación requerida. Según los objetivos iniciales de la investigación.

$$T1 \; / \; T2$$

T1: TIEMPO DE PROCESAMIENTO DE ALGORITMO 1
T2: TIEMPO DE PROCESAMIENTO DE ALGORITMO 2

Tabla 6: Tiempo de Procesamiento entre Red neuronal y ETS-Primaria

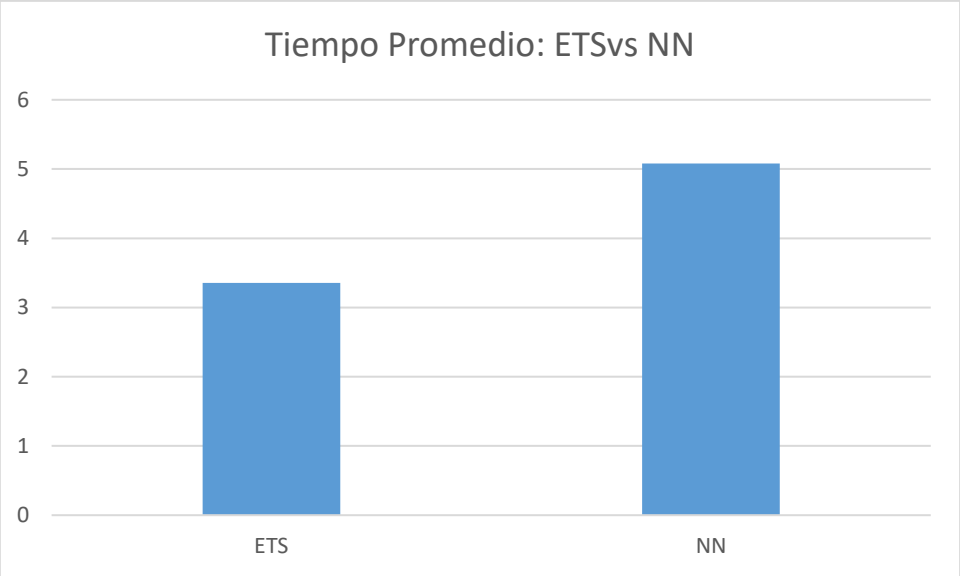
Escenarios Evaluados	ETS	Red neuronal
1	3.26	4.15
2	4.48	5.15
3	2.7	3.85
4	3.15	5.65
5	4.15	4.58
6	3.56	5.85
7	2.87	6.56
8	3.45	5.63
9	4.15	6.23
10	1.78	3.15
Total	3.355	5.08

Fuente: Elaboración Propia

En la Tabla N° 06 podemos apreciar las iteraciones y el tiempo en segundos que demoran los algoritmos para el procesamiento de los datos. El algoritmo de ETS tiene mejor tiempo de procesamiento en cada iteración equivalente a un promedio de 3.355 segundos.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Grafico 3: Tiempo de Procesamiento entre Red Neuronal y ETS-Primaria



Fuente: Elaboración Propia

El gráfico 3 representa el promedio de los algoritmos usados para el procesamiento de los datos, en el cual podemos observar que el algoritmo de redes neuronales es el que mayor tiempo demoró para dicho procesamiento.

Tabla 7: Tiempo de Procesamiento entre Red neuronal y ETS-Secundaria

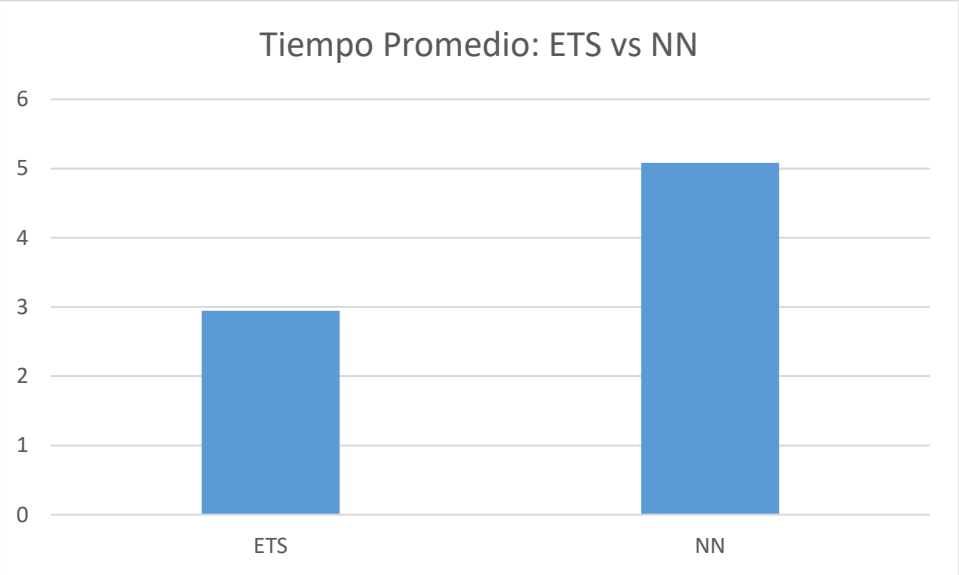
Escenarios Evaluados	ETS	Redneuronal
1	3,03	4,1
2	4,2	5,25
3	2,8	4,05
4	3,01	5,6
5	2,5	4,6
6	3,05	5,8
7	2,09	6,52
8	3,04	5,58
9	4,02	6,2
10	1,7	3,1
Total	2,944	5,08

Fuente: Elaboración Propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En la Tabla N° 07 podemos apreciar las iteraciones y el tiempo en segundos que demoran los algoritmos para el procesamiento de los datos. El algoritmo de ETS tiene mejor tiempo de procesamiento en cada iteración equivalente a un promedio de 2.944 segundos.

Grafico 4: Tiempo de Procesamiento entre Red Neuronal y ETS-Secundaria



Fuente: Elaboración propia

El gráfico 4 representa el promedio de los algoritmos usados para el procesamiento de los datos, en el cual podemos observar que el algoritmo de redes neuronales es el que mayor tiempo demoró para dicho procesamiento.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

C. Tiempo para generar estimación en el sistema

Este indicador mide el tiempo en la solución diseñada, con respecto a la Usabilidad del usuario en el simulador del sistema web para generar un Análisis que obtenga una estimación requerida.

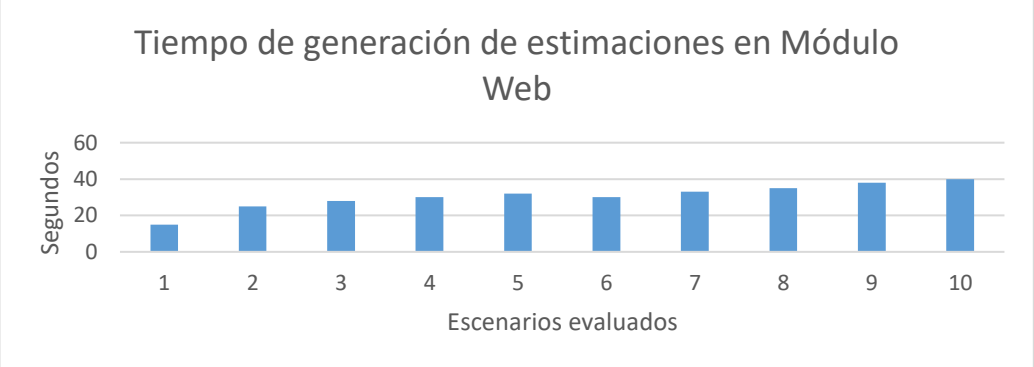
Tabla 8: Tiempo de Procesamiento del Sistema Web-Primaria

Escenarios Evaluados	Sistema Web
1	15
2	25
3	28
4	30
5	32
6	30
7	33
8	35
9	38
10	40
Promedio	30,60 seg.

Fuente: Elaboración Propia

En la Tabla N° 8 se observa que el tiempo promedio de generación de estimaciones en el sistema web para el nivel primario es de 30,60 segundos.

Gráfico 5: Tiempo de generación de pronósticos en Módulo-Primaria



Fuente: Elaboración Propia

El gráfico 5 nos permite observar la variación del tiempo de generación de pronósticos en el Módulo Web para el nivel primario.

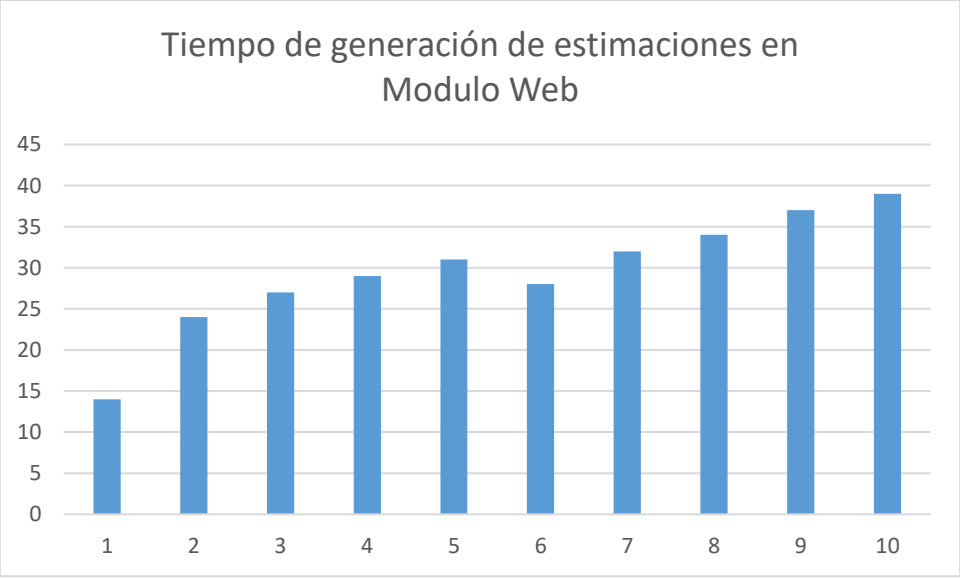
Tabla 9: Tiempo de Procesamiento del Sistema Web-Secundaria

Escenarios Evaluados	Sistema Web
1	14
2	24
3	27
4	29
5	31
6	28
7	32
8	34
9	37
10	39
Promedio	29,5

Fuente: Elaboración propia

En la Tabla N° 9 se observa que el tiempo promedio de generación de estimaciones en el sistema web para el nivel secundario es de 29,5 segundos.

Gráfico 6: Tiempo de generación de pronósticos en Módulo-Secundario



Fuente: Elaboración propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

El gráfico 6 nos permite observar la variación del tiempo de generación de pronósticos en el Módulo Web para el nivel secundario.

4.1. Discusión de resultados

A. Grado de confiabilidad

Con respecto al primer indicador comparando las dos técnicas, es decir Redes Neuronales y ETS. Podemos decir que la red neuronal obtuvo el nivel de confianza más elevado en comparación a ETS, esto se denota en los valores obtenidos al calcular la razón (valor calculado entre el monto real y el monto pronóstico para saber el grado de relación que existe uno con respecto del otro).

B. Tiempo de Procesamiento del Modelo.

En el tiempo de procesamiento al evaluar estas técnicas se obtuvo que con ETS el tiempo promedio de ejecución de 3.355 segundos siendo superior a diferencia de la red neuronal, que tiene 5.08 segundos.

Por otro lado para el nivel secundario el tiempo de procesamiento la técnica ETS obtuvo un tiempo promedio de ejecución de 2.944 segundos siendo superior a diferencia de la red neuronal, que tiene 5.08 segundos.

C. Tiempo para generar estimación en el sistema

Para el último indicador se obtuvo que, en la usabilidad del sistema web, se generó un tiempo promedio de 30.6 segundos para generar una estimación.

Donde para el nivel secundario, se obtuvo que en la usabilidad del sistema, genero un tiempo promedio de 29.5 segundos para generar una estimación.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO V
DESARROLLO DE LA PROPUESTA

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO V: DESARROLLO DE LA PROPUESTA

5.1. Generalidades

Para el desarrollo de la propuesta de investigación se planteó dos metodologías, la metodología CRISP-DM la cual se utilizó para la generación de los modelos aplicando técnicas de minería de datos y la metodología XP, para la construcción del sistema analítico web en el cual se mostraron los resultados del modelo con la data tratada y un simulador con escenarios de comparación de algoritmos de predicción de minería de datos. La justificación de la utilización de las metodologías se expuso en el marco teórico de la presente investigación.

5.2. Metodología

La siguiente investigación consta de dos etapas, la primera que abarca todo lo relacionado al desarrollo de modelos de predicción usando la minería de datos, en esta etapa se contempla todas las fases que se utilizan en la Metodología de desarrollo de modelos de minería de datos (Crisp-DM), desde la comprensión del negocio, datos iniciales, transformación de datos, modelado y aplicación del algoritmo, evaluación de performance. En la segunda etapa se desarrolla la metodología de desarrollo ágil XP, con las fases para el diseño y construcción del sistema web

Se aplica el siguiente marco conceptual para el desarrollo de esta investigación:

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Dado que la investigación tiene como esquema principal, el modelo de minería de datos se ha realizado un cuadro comparativo para la determinación de la metodología que permita resolver esta etapa. Como se puede apreciar en el la tabla 8.

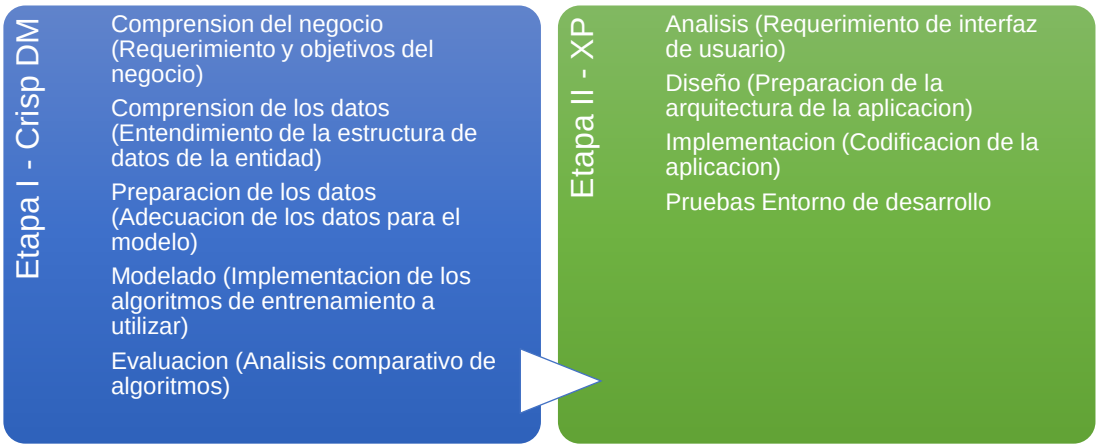
Tabla 10: Metodologías de Desarrollo de Modelo de Minería de Datos

	CRISP-DM	SEMMA
Libre elección de herramientas	SI	NO
Cantidad de fases	6	5
Todas las fases pueden relacionar	SI	NO
Procesos de Inteligencia de Negocios	SI	NO
Comercial – Licencias – Privativa	NO	SI
Técnicas de ETL	SI	SI

Fuente: (Flores, 2009)

Se establece usar CRISP-DM, por ser una metodología flexible en cuanto a herramientas, además que integra el proceso de comprensión de negocio (Gestión del proyecto por objetivos empresariales), en cuanto la metodología SEMMA es una buena alternativa siempre y cuando se use en proyecto con tecnologías SAS.

Figura 9: Etapas de Desarrollo



Fuente: Elaboración propia

5.2.1. Metodología CRISP DM Minería de Datos

5.2.1.1. Comprensión del negocio

La Ugel Chiclayo es una entidad gubernamental. Donde cada fin del año escolar se hace de manera manual el consolidado de todos los alumnos de cada institución educativa las que conforman la Ugel Chiclayo.

Donde toda esa información que se pasan a un archivo Excel y de esa manera tener registrados el total de alumnos matriculados en cada respectiva institución educativa.

Tabla 11: Periodo - Matriculados

Año	Matriculados
2006	154466
2007	375200
2008	148647
2009	79156
2010	78009
2011	77918
2012	74050
2013	104116
2014	76860
2015	77483

Fuente: Elaboración Propia

b. Necesidades y Expectativas

b.1.Búsqueda de la mejora en las predicciones con respecto a los alumnos matriculados de una institución educativa en un periodo determinado.

b.2. Implementar una nueva y mejor técnica en cuanto al proceso predictivo.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

c. Objetivos de Negocio

c.1 Analizar tendencias de predicción con respecto a los alumnos matriculados de un determinada Institución.

c.2 Realizar pronósticos de forma anual, con base en un nivel de confianza previamente definido en un periodo determinado.

d. Criterios de Éxito

d.1 Confiabilidad de los pronósticos realizados en un determinado periodo.

d.2 Facilidad de acceso con respecto al aplicativo web.

e. Evaluación de la situación

e.1 Se cuenta con la base de datos de alumnos matriculados en la región de Lambayeque desde el año 2006. Esta información es utilizada como fuente principal para la creación del modelo de series de tiempo.

f. Requerimientos

f.1 El sistema permitirá la generación de reportes para la visualización de las predicciones de alumnos que podrían desertar en los próximos años.

f.2 Visualizar la comparación de modelos predictivos y utilizando el mejor para beneficios de la institución educativa.

g. Restricciones

g.1 Se requiere la base de datos de todos los alumnos matriculados desde hace 9 años de antigüedad como mínimo para el entrenamiento y testeo del modelo.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

g.2 De la información obtenida, los datos deben estar libre de errores y valores en valores en blanco.

h. Determinación de los Objetivos de minería de datos

h.1 Objetivos del Proyecto

h.1.1 Generar un modelo de series de tiempo, que arroje predicciones con un alto grado de confianza en un tiempo determinado.

h.2.2 Entrenar el modelo para su mejor eficiencia.

h.3.3 Testear el modelo para el resultado.

h.2 Criterios de éxito del proyecto

h.2.1 Confiabilidad del modelo diseñado e implementado.

h.2.2 Optimización del tiempo para la generación de reportes.

5.2.1.2. Comprensión de los datos

A. Recolección de los Datos del Negocio Iniciales

A.1 Proceso de Adquisición

Los datos obtenidos corresponden a los alumnos matriculados en la Ugel de Chiclayo de forma anual y por colegios.
No se realizará una transformación de datos ya que la información con la que se cuenta es real; dichos datos son utilizados como ingreso para el entrenamiento del modelo.

A.2. Selección de las Variables a utilizar

Para la creación del modelo con series de tiempo, los atributos utilizados son identificados de la siguiente manera: El atributo codigocolegio se denota como el código de cada colegio, el atributo Dato01h es la cantidad de hombres matriculados en 1 grado de primaria, Dato01M es la cantidad de mujeres matriculados en 1 grado de primaria, el atributo Dato02h es la cantidad de hombres

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

matriculados en 2 grado de primaria, Dato02M es la cantidad de mujeres matriculados en 2 grado de primaria, el atributo Dato03h es la cantidad de hombres matriculados en 3 grado de primaria, Dato03M es la cantidad de mujeres matriculados en 3 grado de primaria, el atributo Dato4h es la cantidad de hombres matriculados en 4 grado de primaria, Dato04M es la cantidad de mujeres matriculados en 4 grado de primaria, el atributo Dato5h es la cantidad de hombres matriculados en 5 grado de primaria, Dato05M es la cantidad de mujeres matriculados en 5 grado de primaria, el atributo Dato6h es la cantidad de hombres matriculados en 6 grado de primaria, Dato06M es la cantidad de mujeres matriculados en 6 grado de primaria, el atributo CODOOII es el código de la ugel de Chiclayo CODOOII=140001, el atributo NIV_MOD es el código del nivel Bo=primaria, Fo=secundaria, el atributo AREA es para identificar 1=rural, 2=urbano, como se detalla en la siguiente imagen.

Figura 10: Datos de alumnos Matriculados

	COD_MO	CodigoCole	DATO	DATO01	DA	DA	DA	DA	DA	DA	DA	DA	DA	DAT	DA	DA	DA	DA	DA	DATO	DATO	DATO10	DAT	CODOOII	NIV	AREA
16	0262196		0	0	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
17	0262196		0	0	0	0	159	0	99	0	115	0	99	0	0	0	0	0	0	0	0	0	0	010001	F0	1
18	0262196		0	0	0	0	137	0	90	0	106	0	95	0	0	0	0	0	0	0	0	0	0	010001	F0	1
19	0262196		0	0	0	0	22	0	9	0	9	0	4	0	0	0	0	0	0	0	0	0	0	010001	F0	1
20	0262196		0	0	0	0	4	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	010001	F0	1
21	0262196		0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	010001	F0	1
22	0262196		0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
23	0262196		0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
24	0262196		0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
25	0262196		0	0	0	0	0	0	0	0	24	58	11	2	2	0	0	0	0	0	0	0	0	010001	F0	1
26	0262196		0	0	101	0	163	0	100	0	116	0	100	0	0	0	0	0	0	0	0	0	0	010001	F0	1
27	0262196		0	0	101	0	163	0	100	0	116	0	100	0	0	0	0	0	0	0	0	0	0	010001	F0	1
342	0678581		0	66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
343	0678581		0	0	0	66	0	56	0	54	0	57	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
344	0678581		0	0	0	66	0	54	0	52	0	54	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
345	0678581		0	0	0	0	0	2	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
346	0678581		0	0	0	0	0	0	0	0	0	34	31	0	1	0	0	0	0	0	0	0	0	010001	F0	1
347	0678581		0	66	0	66	0	56	0	54	0	57	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
348	0678581		0	66	0	66	0	56	0	54	0	57	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1
349	1302017		0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	010001	F0	1

Fuente: Base de datos en excel “Mat2x00”

Tratamiento de nulos (Directo)

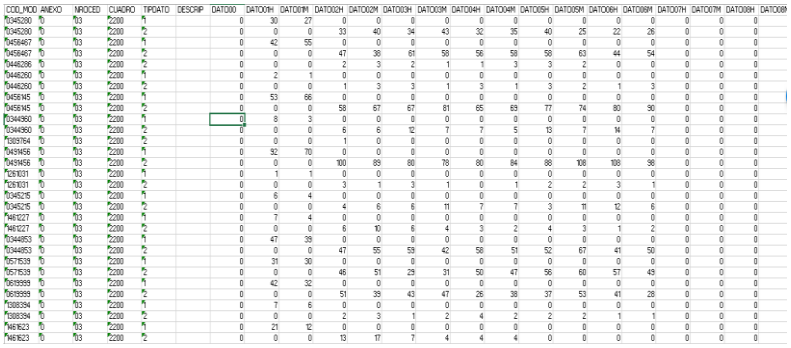
Los valores nulos se pueden dar por las siguientes razones:

- No se pudo capturar cuantos alumnos están matriculados.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

En el nulo directo, aun así cuando en el registro de alumnos matriculados en cada aula.

Figura 11: Tratamiento de Datos Nulos



COD_MOD	ANEXO	NROCED	CUADRO	TIPOATO	DESCRIP	DATO00	DATO01H	DATO01M	DATO02H	DATO02M	DATO03H	DATO03M	DATO04H	DATO04M	DATO05H	DATO05M	DATO06H	DATO06M	DATO07H	DATO07M	DATO08H	DATO08M
0345280	0	03	2200	1		0	30	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0345280	0	03	2200	2		0	0	0	33	40	34	43	32	35	40	25	22	26	0	0	0	0
0456467	0	03	2200	1		0	42	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0456467	0	03	2200	2		0	0	0	47	38	61	59	56	59	58	63	44	54	0	0	0	0
0446286	0	03	2200	2		0	0	0	2	3	2	1	1	3	3	2	0	0	0	0	0	0
0446260	0	03	2200	1		0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0446260	0	03	2200	2		0	0	0	1	3	3	1	3	1	3	2	1	3	0	0	0	0
0456745	0	03	2200	1		0	53	66	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0456745	0	03	2200	2		0	0	0	58	67	67	81	65	63	77	74	80	50	0	0	0	0
0344960	0	03	2200	1		0	8	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0344960	0	03	2200	2		0	0	0	6	6	12	7	7	5	13	7	14	7	0	0	0	0
0389764	0	03	2200	2		0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0439456	0	03	2200	1		0	92	70	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0439456	0	03	2200	2		0	0	0	100	89	88	76	80	84	88	108	100	98	0	0	0	0
0261031	0	03	2200	1		0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0261031	0	03	2200	2		0	0	0	3	1	3	1	0	1	2	2	3	1	0	0	0	0
0345215	0	03	2200	1		0	6	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0345215	0	03	2200	2		0	0	0	4	6	6	11	7	7	3	11	12	6	0	0	0	0
0481227	0	03	2200	1		0	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0481227	0	03	2200	2		0	0	0	6	10	8	4	3	2	4	3	1	2	0	0	0	0
0344853	0	03	2200	1		0	47	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0344853	0	03	2200	2		0	0	0	47	55	59	42	58	51	52	67	41	50	0	0	0	0
0571539	0	03	2200	1		0	31	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0571539	0	03	2200	2		0	0	0	46	51	29	31	50	47	56	60	57	49	0	0	0	0
0619999	0	03	2200	1		0	42	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0619999	0	03	2200	2		0	0	0	51	39	43	47	26	38	37	53	41	28	0	0	0	0
0388394	0	03	2200	1		0	7	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0388394	0	03	2200	2		0	0	0	2	3	1	2	4	2	2	2	1	1	0	0	0	0
0481623	0	03	2200	1		0	21	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0481623	0	03	2200	2		0	0	0	13	17	7	4	4	4	0	0	0	0	0	0	0	0

Fuente: Elaboración propia

Luego de limpiar los datos nulos del archivo de Microsoft Excel “MAT2X00.xls”, donde luego se pasara a una base de datos “Deserción”. Así como lo visualiza en la siguiente imagen.

Figura 12: Datos Sin datos Nulos

Año	CodigoColegio	COD_MOD	DATO01H	DATO01M	DATO02H	DATO02M	DATO03H	DATO03M	DATO04H	DATO04M	DATO05H	DATO05M	DATO06H	DATO06M	Nivel
2006	277041	0217026	6	3	18	10	14	14	22	6	22	6	12	12	BO
2006	275706	0344820	51	55	86	118	116	84	82	146	92	150	86	110	BO
2006	275711	0344838	81	89	146	136	126	132	140	118	168	150	110	120	BO
2006	673673	0344846	71	62	140	132	154	216	186	192	194	170	190	228	BO
2006	280109	0344853	47	39	110	132	132	92	122	114	108	134	84	100	BO
2006	278074	0344879	27	29	76	40	68	46	100	66	74	102	48	78	BO
2006	278578	0344887	62	85	166	150	176	190	192	194	198	210	142	188	BO
2006	278597	0344903	40	35	84	66	96	78	72	72	66	70	50	62	BO
2006	278601	0344911	88	76	208	242	202	230	220	198	236	224	352	204	BO
2006	280034	0344929	25	21	92	82	56	56	66	64	92	92	72	90	BO
2006	280114	0344945	22	27	68	82	86	92	84	88	80	72	66	46	BO
2006	280128	0344960	8	3	22	20	32	16	14	10	26	14	28	14	BO
2006	280500	0344978	31	29	90	88	52	100	120	78	92	96	94	72	BO
2006	280519	0344986	42	40	118	96	76	104	94	130	94	104	64	82	BO
2006	280604	0345009	52	52	106	82	124	108	130	114	86	98	106	90	BO
2006	275725	0345025	101	97	252	224	204	332	258	272	166	326	156	368	BO
2006	275532	0345033	23	19	38	26	60	36	26	22	46	38	40	36	BO
2006	275730	0345041	38	41	128	116	108	134	116	126	96	140	150	104	BO
2006	275749	0345066	35	38	88	76	70	98	80	88	56	100	76	86	BO
2006	279436	0345082	25	22	48	44	34	40	62	76	72	58	72	56	BO
2006	279422	0345108	2	2	0	16	2	4	12	8	20	12	12	8	BO
2006	279403	0345116	3	5	8	14	12	8	8	6	2	18	18	6	BO
2006	279568	0345132	13	6	14	20	16	18	18	22	22	18	18	14	BO
2006	279573	0345140	52	37	58	78	48	60	64	38	48	58	36	54	BO
2006	279587	0345157	15	10	16	14	20	28	18	22	32	28	14	30	BO

Fuente: Elaboración propia

A.3. Datos y métodos de captura

Los datos han sido extraídos de la base de datos que están almacenados en Microsoft Excel. Para luego pasarlos a la herramienta Spss para luego pasarlo al formato requerido como se muestra en la imagen.

Figura 13: Tratamiento de Datos

Tablas personalizadas

NIV_MOD B0

		DAT001H	DAT001M	DAT002H	DAT002M	DAT003H	DAT003M	DAT004H	DAT004M	DAT005H	DAT005M	DAT006H	DAT006M
		Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma
COD_MOD	0217026	12	7	35	15	21	21	33	9	33	9	18	18
	0344820	129	132	335	183	174	126	123	220	138	225	129	165
	0344838	162	178	398	207	189	198	210	177	252	225	165	180
	0344846	142	124	339	202	231	324	279	288	291	255	285	342
	0344853	94	78	253	229	201	145	184	171	162	201	126	150
	0344879	49	63	168	62	102	69	150	99	111	153	72	117
	0344887	118	168	497	239	264	285	288	291	297	315	213	282
	0344895	38	48	146	97	93	118	115	130	105	93	123	117
	0344903	80	70	182	110	152	119	108	108	100	106	75	93
	0344911	175	196	573	363	303	345	330	297	354	336	528	306
	0344929	61	116	162	123	84	84	99	96	138	138	108	135
	0344945	44	57	165	129	133	140	126	133	120	108	99	69
	0344960	28	18	55	30	48	24	21	15	39	21	42	21
	0344978	62	58	193	135	81	151	181	117	138	144	141	108
	0344986	97	97	260	164	117	158	143	196	141	156	96	123
	0345009	102	98	273	139	186	162	195	171	129	147	159	135

Fuente: “Spss”

Donde luego de haber homogenizado la data se exporta a un archivo Excel para cada uno de los años como se muestra en la imagen:

Figura 14: Datos Tratados

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Año	CodigoColegio	COD_MOD	DAT001H	DAT001M	DAT002H	DAT002M	DAT003H	DAT003M	DAT004H	DAT004M	DAT005H	DAT005M	DAT006H	DAT006M	Nivel
2	2006	277041	0217026	6	3	18	10	14	14	22	6	22	6	12	12	BO
3	2006	275706	0344820	51	55	86	118	116	84	82	146	92	150	86	110	BO
4	2006	275711	0344838	81	89	146	136	126	132	140	118	168	150	110	120	BO
5	2006	673673	0344846	71	62	140	132	154	216	186	192	194	170	190	228	BO
6	2006	280109	0344853	47	39	110	132	132	92	122	114	108	134	84	100	BO
7	2006	278074	0344879	27	29	76	40	68	46	100	66	74	102	48	78	BO
8	2006	278578	0344887	62	85	166	150	176	190	192	194	198	210	142	188	BO
9	2006	278597	0344903	40	35	84	66	96	78	72	72	66	70	50	62	BO
10	2006	278601	0344911	88	76	208	242	202	230	220	198	236	224	352	204	BO
11	2006	280034	0344929	25	21	92	82	56	56	66	64	92	92	72	90	BO
12	2006	280114	0344945	22	27	68	82	86	92	84	88	80	72	66	46	BO
13	2006	280128	0344960	8	3	22	20	32	16	14	10	26	14	28	14	BO
14	2006	280500	0344978	31	29	90	88	52	100	120	78	92	96	94	72	BO
15	2006	280519	0344986	42	40	118	96	76	104	94	130	94	104	64	82	BO
16	2006	280604	0345009	52	52	106	82	124	108	130	114	86	98	106	90	BO
17	2006	275725	0345025	101	97	252	224	204	332	258	272	166	326	156	368	BO
18	2006	275532	0345033	23	19	38	26	60	36	26	22	46	38	40	36	BO
19	2006	275730	0345041	38	41	128	116	108	134	116	126	96	140	150	104	BO
20	2006	275749	0345066	35	38	88	76	70	98	80	88	56	100	76	86	BO
21	2006	279436	0345082	25	22	48	44	34	40	62	76	72	58	72	56	BO

Fuente: Base de datos en Excel “Mat2x00”

Después de haber pasado los datos a un formato homogéneo se realizó la migración manual de documentos ofimáticos a la base de datos “Deserción”, a partir de estas tablas se procede a realizar el modelo predictivo.

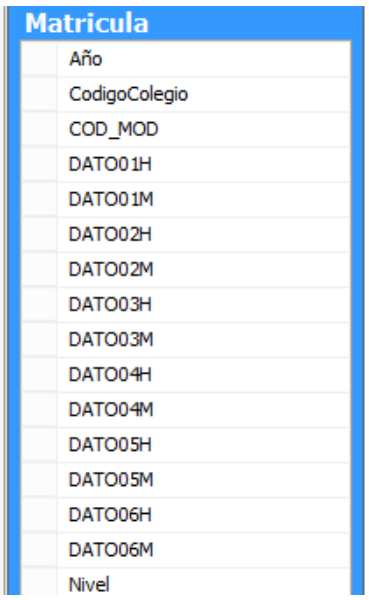
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

A.4. Exploración de Datos

La construcción del modelo de predicción se desarrolla con información obtenida desde el año 2006 hasta el año 2015. Estos datos son los que ingresan en una pequeña base de datos obtenida por la migración de datos en repositorios ofimáticos a la base de datos “Deserción” en el gestor SQL SERVER 2008 para que realice el entrenamiento del modelo; de los cuales se utiliza el 70% para el entrenamiento y el 30% para las pruebas de predicciones.

Al realizar este proceso de aprendizaje en el modelo se obtiene un valor aproximado que medirá el rendimiento del modelo mostrando el porcentaje de error, el cuan deberá ser mínimo para demostrar que el modelo está bien creado con un alto grado de certeza.

Figura 15: Diagrama E-R Esquema Matriculas



Fuente: Elaboración propia

Después del proceso de aprendizaje del modelo se obtiene un valor aproximado que medirá el rendimiento del modelo mostrando el porcentaje de error, el cual deberá ser mínimo para demostrar que el modelo esta creado con un grado de certeza muy alto.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Tabla 12: Alumnos matriculados 2006-2015

Año	Total_matricula
2006	154466
2007	375200
2008	148647
2009	79156
2010	78009
2011	77918
2012	74050
2013	104116
2014	76860
2015	77483

Fuente: Elaboración propia

5.2.1.3. Preparación de los datos

A. Datos Seleccionados

De la base de datos obtenida, se obtienen diferentes tipos de información con respecto a los alumnos matriculados, lo cual son datos relevantes, para ello, se ha realizado un análisis de la data con los atributos a utilizar para el correcto funcionamiento del modelo. Debe considerarse además que se ha analizado y utilizado el campo Nivel, para el proceso de limpieza de datos.

Figura 16: Scripts SQL para análisis de data

```
Create view Primaria
select Año, sum(DATO01H+DATO01M)as 'Primero',sum(DATO02H+DATO02M)as 'Segundo',
sum(DATO03H+DATO03M)as 'Tercero',
sum(DATO04H+DATO04M)as 'Cuarto',
sum(DATO05H+DATO05M)as 'Quinto',
sum(DATO06H+DATO06M)as 'Sexto',
sum(DATO01H+DATO01M+DATO02H+DATO02M+DATO03H+DATO03M+DATO04H+DATO04M+DATO05H+DATO05M+DATO06H+DATO06M)as 'Total',
CodigoColegio
from Matricula where nivel='BO'
group by Año,CodigoColegio
```

Fuente: Elaboración propia

B. Estructuración de los datos

Para la creación del modelo con series de tiempo, los atributos utilizados son identificados de la siguiente manera: al atributo Año, Primero, Segundo, Tercero, Cuarto, Quinto, Sexto, Total que se denota a la cantidad total de alumnos ya que representa el objetivo a predecir. Como se demuestra en nuestra imagen.

Figura 17: Data para análisis

	Año	Primero	Segundo	Tercero	Cuarto	Quinto	Sexto	Total	CodigoColegio
1	2006	4	12	14	12	10	2	54	001976
2	2007	30	40	32	18	15	9	144	001976
3	2008	10	16	14	12	10	6	68	001976
4	2009	12	9	3	5	6	4	39	001976
5	2010	7	9	9	3	3	5	36	001976
6	2011	2	11	9	7	4	3	36	001976
7	2012	0	3	8	6	7	4	28	001976
8	2013	2	2	5	6	9	5	29	001976
9	2014	2	3	2	4	7	9	27	001976
10	2015	5	2	2	2	5	9	25	001976
11	2009	18	25	17	14	18	0	92	008082

Fuente: Elaboración propia

5.2.1.4. Modelado

En la investigación se propone construir un modelo de minería de datos utilizando técnicas de pronósticos, a continuación, se presenta la tabla que se realizó para la selección de las técnicas adecuadas.

Tabla 13: Evaluación de las técnicas de minería de datos

TÉCNICA DE MINERÍA DE DATOS	DESCRIPCIÓN DE LA TÉCNICA	ALGORITMOS	¿ES ADECUADO PARA LA INVESTIGACIÓN?
REGRESIÓN	Modelos de 2 variables	Redes Neuronales, ETS	SI
ASOCIACION	Hechos en común para determinado grupo de datos múltiples variables	A priori FP-Growth Éclat	NO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CLASIFICACION AD HOC	Basado en reglas por construcciones lógicas múltiples variables	Árbol de decisiones, Redes Bayesianas	NO

Fuente: Elaboración propia

Para lo cual se han establecido los siguientes criterios de evaluación de los Algoritmos a utilizar.

Tabla 14: Modelos de Minería de Datos

	ETS	HOLT	RED NEURONAL AUTO REGRESIVA
Evaluación fundamento teórico			
Modelo parametrizado	**X**	**X**	**----**
Datos estacionales	**X**	**X**	**X**
Método estadístico	**X**	**X**	**----**
Capacidad iterativa (Aprendizaje)	**-----**	**-----**	**X**
Cantidad de datos de la serie	25	28	3
Evaluación fundamento computacional			
Procesamiento CPU	**Mínimo**	**Mínimo**	**Medio**
Consumo RAM	**Mínimo**	**Mínimo**	**Medio**
Tiempo computacional	**Mínimo**	**Mínimo**	**Medio**
Evaluación fundamento objetivo del modelo			
Confiabilidad de precisión pronostico	**Después de pruebas**	**Después de pruebas**	**Después de pruebas**
Confiabilidad de precisión consistencias	**Después de pruebas**	**Después de pruebas**	**Después de pruebas**

Fuente: Elaboración propia

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Se ha considerado usar ETS y REDES NEURONALES, donde ETS se utilizó por requerir la cantidad de datos necesarios con la que se dispone en el histórico de cada colegio, y la red neuronal auto regresiva se utilizó por la naturaleza de la investigación donde se utiliza series de tiempo como refiere (Vílchez García, 2010) , sin embargo para este caso debido a la cantidad de datos se cuenta no es factible emplear el algoritmo Holt. Después de la pruebas de laboratorio como se muestra en el Anexo 3 se determinó que el tiempo de procesamiento es mínimo

5.2.1.4.1. Modelo A

5.2.1.4.1.1. Descripción del Modelo A

Nnetar es una Red neuronal auto regresivo, el modelo es de tipo regresión, la cual analiza el comportamiento de múltiples variables para determinar un estado objetivo.

Figura 18: Algoritmo R - Nnetar

```
nnetar <- function(x, p, P=1, size, repeats=20, lambda=NULL)
{
  # Transform data
  if(!is.null(lambda))
    xx <- BoxCox(x,lambda)
  else
    xx <- x

  # Scale data
  scale <- max(abs(xx),na.rm=TRUE)
  xx <- xx/scale
  # Set up lagged matrix
  n <- length(xx)
  xx <- as.ts(xx)
  m <- frequency(xx)
  if(m==1)
  {
    if(missing(p))
      p <- max(length(ar(na.interp(xx))$ar),1)
    lags <- 1:p
    P <- 0
  }
  else
  {
    if(missing(p))
    {
      x.sa <- seasadj(stl(na.interp(xx),s.window=7))
      p <- max(length(ar(x.sa)$ar),1)
    }
    if(P > 0)
      lags <- sort(unique(c(1:p,m*(1:P))))
  }
  if(missing(size))
    size <- round((p+P+1)/2)
  maxlag <- max(lags)
  nlag <- length(lags)
  y <- xx[-(1:maxlag)]
```

Fuente: R Project 3.2.2

En el caso de la investigación, el único valor que se ingresa está dado por un vector numérico de series de tiempo.

El análisis de la serie entonces por una red neuronal debe tratarse con un método previo, que es la teoría de ventanas, se trata de un algoritmo que expande y genera atributos (columnas) a partir de los datos iniciales del vector, por lo que al generar dichos atributos se trata de explicar la relación de estos a partir de un modelo regresivo.

5.2.1.4.1.2. Evaluación del Modelo A

En R aplicamos el algoritmo al histórico del colegio, el modelo realiza el entrenamiento de la serie donde determina de manera automática e interpretativa los valores de componentes de la serie de tiempo.

Tabla 15: Datos algoritmo Red Neuronal

Año	V Original	V-1	V-2	V-3
2010	45	¿	¿	¿
2011	65	45	¿	¿
2012	55	65	45	¿
2013	75	55	65	45
2014	89	75	55	65
2015	13	89	75	55
2016(Objetivo)	X	¿	¿	¿

Fuente: Elaboración Propia

En la tabla anterior la tomamos con un ejemplo para mostrar que el formato a analizar por la red neuronal trata de explicar el fenómeno obteniendo para cada año, además se puede apreciar cómo se distribuye según la información que contiene el vector original.

A continuación, se presenta el Algoritmo Nnetar en líneas de código.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Figura 19: Aplicación del Algoritmo Nnetar

```
fitnnseg <- nnetar (seg,2,P=1,2,repeats=20)
pronnseg<- forecast(fitnnseg, h=1)

fitnnter <- nnetar (ter,2,P=1,2,repeats=20)
pronnter<- forecast(fitnnter, h=1)

fitnncua <- nnetar (cua,2,P=1,2,repeats=20)
pronncua<- forecast(fitnncua, h=1)

fitnnqui <- nnetar (qui,2,P=1,2,repeats=20)
pronnqui<- forecast(fitnnqui, h=1)

fitnnsex <- nnetar (sex,2,P=1,2,repeats=20)
pronnsex<- forecast(fitnnsex, h=1)


pronnseg<-round(pronnseg$mean[1],0)
pronnter<-round(pronnter$mean[1],0)
pronncua<-round(pronncua$mean[1],0)
pronnqui<-round(pronnqui$mean[1],0)
pronnsex<-round(pronnsex$mean[1],0)
```

Fuente: Elaboración propia

Después de Implementar el algoritmo Nnetar, en la siguiente imagen se puede apreciar el funcionamiento del algoritmo.

Figura 20: Nnetar

IdLaboratorio	IdTipo	Anio	IdColegio	Rds	Rdt	Rdc	Rdq	Rdse	ProDSnn	ProDTnn	ProDCnn	ProDQnn	ProDSenn	ProDSets	ProDTets	ProDCets	ProDQets	ProDSeets
1	1	2015	277041	4	3	9	3	-3	2	2	2	-6	1	7	-2	-6	-2	-1
2	1	2015	275706	8	4	12	3	3	4	-9	-8	5	-6	48	-31	-49	-10	-22
3	1	2015	275711	1	4	2	-23	0	2	-26	39	146	-125	77	-59	-53	-26	-15
4	1	2015	673673	7	-4	-7	4	-6	18	-12	-8	1	-11	94	-60	-67	-16	-39
5	1	2015	280109	12	12	5	13	0	7	2	-6	-1	-8	40	-12	-49	-21	-14
6	1	2015	278074	9	-11	3	-7	-5	4	-10	-5	0	-1	23	-10	-29	-3	-13
7	1	2015	278578	14	-25	9	-4	-2	6	0	-10	-56	5	98	-44	-56	-36	-20
8	1	2015	278597	12	-1	0	8	0	4	-9	-16	-8	-77	28	-28	-39	-17	-16
9	1	2015	278601	31	2	12	24	0	20	-14	-12	-13	-3	108	-42	-89	-38	-22
10	1	2015	280034	1	-9	5	6	-1	-6	-5	-2	2	3	35	-19	-30	-7	-9

Fuente: Elaboración propia.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.2.1.4.2. Modelo B

5.2.1.4.2.1. Descripción del Modelo B

Es una regla de la técnica general para suavizar los datos de series de tiempo, sobre todo para aplicar de forma recursiva hasta tres filtros de paso bajo con funciones de la ventana exponenciales.

Figura 21: Algoritmo ETS

```
ets <- function(y, model="ZZZ", damped=NULL,
    alpha=NULL, beta=NULL, gamma=NULL, phi=NULL, additive.only=FALSE, lambda=NULL,
    lower=c(rep(0.0001,3), 0.8), upper=c(rep(0.9999,3),0.98),
    opt.crit=c("lik","amse","mse","sigma","mae"), nmse=3, bounds=c("both","usual","admissible"),
    ic=c("aicc","aic","bic"),restrict=TRUE, allow.multiplicative.trend=FALSE,
    use.initial.values=FALSE, ...)
{
  #dataname <- substitute(y)
  opt.crit <- match.arg(opt.crit)
  bounds <- match.arg(bounds)
  ic <- match.arg(ic)

  #if(max(y,na.rm=TRUE) > 1e6)
  #    warning("Very large numbers which may cause numerical problems. Try scaling the data first")

  if(any(class(y) %in% c("data.frame","list","matrix","mts")))
    stop("y should be a univariate time series")
  y <- as.ts(y)

  # Check if data is constant
  if (is.constant(y))
    return(ses(y, alpha=0.99999, initial='simple')$model)

  # Remove missing values near ends
  ny <- length(y)
  y <- na.contiguous(y)
  if(ny != length(y))
    warning("Missing values encountered. Using longest contiguous portion of time series")
```

Fuente: (Hyndman R. , 2015)

A continuación, se presenta el Algoritmo ETS en líneas de código.

Figura 22: Aplicación del Algoritmo ETS

```
fitetsseg <- ets(seg)
proetsseg<- forecast(fitetsseg, h=1)

fitetster <- ets(ter)
proetster<- forecast(fitetster, h=1)

fitetscua <- ets(cua)
proetscua<- forecast(fitetscua, h=1)

fitetsqui <- ets(qui)
proetsqui<- forecast(fitetsqui, h=1)

fitetssex <- ets(sex)
proetssex<- forecast(fitetssex, h=1)


proetsseg<-round(proetsseg$mean[1],0)
proetster<-round(proetster$mean[1],0)
proetscua<-round(proetscua$mean[1],0)
proetsqui<-round(proetsqui$mean[1],0)
proetssex<-round(proetssex$mean[1],0)
```

Fuente: Elaboración propia

Después de Implementar el algoritmo ETS, en la siguiente imagen se puede apreciar el funcionamiento del algoritmo.

Figura 22: ETS

IdLaboratorio	IdTipo	Anio	IdColegio	Rds	Rdt	Rdc	Rdq	Rdse	ProDSnn	ProDTnn	ProDCnn	ProDQnn	ProDSenn	ProDSets	ProDTets	ProDCets	ProDQets	ProDSeets
1	1	2015	277041	4	3	9	3	-3	2	2	2	-6	1	7	-2	-6	-2	-1
2	1	2015	275706	8	4	12	3	3	4	-9	-8	5	-6	48	-31	-49	-10	-22
3	1	2015	275711	1	4	2	-23	0	2	-26	39	146	-125	77	-59	-53	-26	-15
4	1	2015	673673	7	-4	-7	4	-6	18	-12	-8	1	-11	94	-60	-67	-16	-39
5	1	2015	280109	12	12	5	13	0	7	2	-6	-1	-8	40	-12	-49	-21	-14
6	1	2015	278074	9	-11	3	-7	-5	4	-10	-5	0	-1	23	-10	-29	-3	-13
7	1	2015	278578	14	-25	9	-4	-2	6	0	-10	-56	5	98	-44	-56	-36	-20
8	1	2015	278597	12	-1	0	8	0	4	-9	-16	-8	-77	28	-28	-39	-17	-16
9	1	2015	278601	31	2	12	24	0	20	-14	-12	-13	-3	108	-42	-89	-38	-22
10	1	2015	280034	1	-9	5	6	-1	-6	-5	-2	2	3	35	-19	-30	-7	-9

Fuente: Elaboración propia.

5.2.1.4.2.2. Evaluación del Modelo B

En R aplicamos el algoritmo al histórico del colegio, el modelo realiza el entrenamiento de la serie donde determina de manera automática e interpretativa los valores de componentes de la serie de tiempo.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

5.2.1.5. Etapa II – Metodología XP para el desarrollo de aplicación web

a) Planificación del Proyecto

Tabla 16: Prioridad y Dificultad de Historia de Usuario

HISTORIA DE USUARIO	PRIORIDAD	N° ITERACIONES
1. CONSULTAR Y GENERAR REPORTES	ALTA	3
2. GENERAR PROYECCIONES Y ESTIMACIONES.	ALTA	3
3. GESTION DE USUARIOS.	MEDIA	2
4. GESTION DE REPORTES	MEDIA	2

Fuente: Extraído de la Metodología XP

La prioridad está definido por el aspecto del sistema, es decir que está en función principal por las historias de usuarios.

Historia de usuario detallado

Tabla 17: Requerimiento 01

Historia de Usuario	
Número: 1	**Usuario: Estadisticos,Especialistas,estudiantes,invitados**
Nombre historia: CONSULTAR Y GENERAR REPORTES	
Prioridad en negocio: Alta	**Riesgo en desarrollo: Baja**
Entrevistado:Estadistico,Usuario Funcional	
Descripción:Podrán acceder al módulo de monitoreo de información anual.	
Observaciones:	

Fuente: Elaboración Propia

Tabla 18: Requerimiento 02

Historia de Usuario	
Número: 2	**Usuario:Analistas de datos**
Nombre historia: GENERAR PROYECCIONES Y ESTIMACIONES	
Prioridad en negocio: Alta	**Riesgo en desarrollo: Baja**

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Entrevistado: Analista de datos
Descripción: **El analista de datos podrá entrar en el módulo de proyecciones y estimaciones donde podrán simular con los datos cualquier escenario posible que le permita el sistema de análisis, puede visualizar el modelo por defecto o generar nuevos valores a partir de simulaciones.**
Observaciones:

Fuente: Elaboración Propia

Tabla 19: Requerimiento 03

Historia de Usuario	
Número: 3	**Usuario: Administrador del Sistema**
Nombre historia: GESTIÓN DE USUARIOS	
Prioridad en negocio: **Alta**	**Riesgo en desarrollo:** **Baja**
Entrevistado:	
Descripción: **El sistema contará con 2 niveles de usuario: Administrador y Estadísticos. Cada uno de ellos tendrá restricciones en el sistema.** **Administrador: Acceso a todos los módulos del sistema.** **Estadísticos: Acceso a la visualización de reportes del modelo, que son los resultados de las predicciones.** **El sistema debe permitir, visualizar y estructurar nuevos reportes.**	
Observaciones:	

Fuente: Elaboración Propia

B) Diseño

Base de Datos Relacional

El sistema está diseñado para cumplir dos propósitos, la captura de los datos que viene a ser la migración de los documentos ofimáticos en función al consolidado de Matriculados, siendo este la mínima unidad representativa de tiempo registrado, por lo tanto, el sistema contempla esta captura de datos y el almacenamiento de información por parte de la ejecución del modelo, así como los datos administrativos del sistema.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

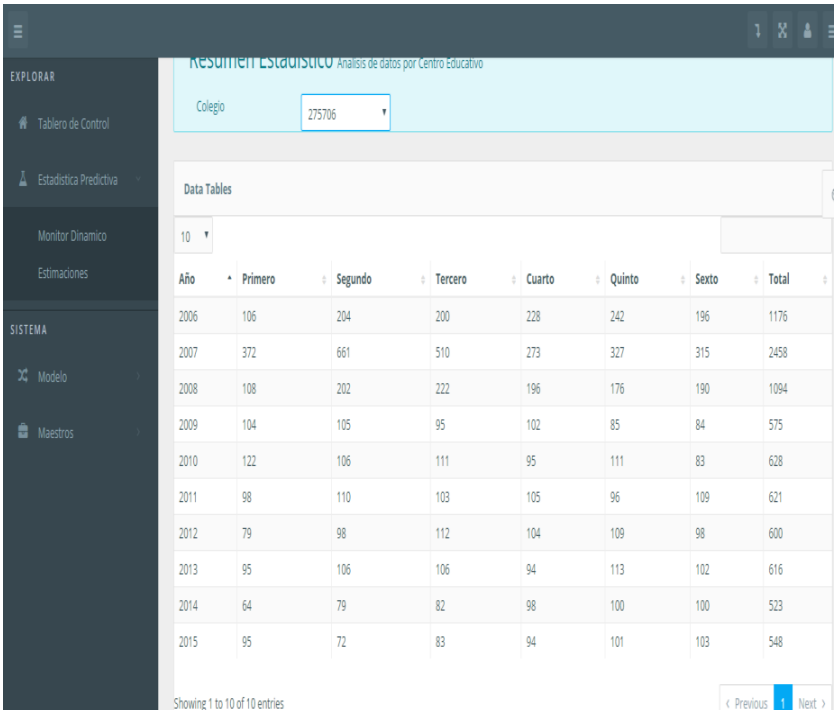
C) Interfaz web de simulaciones

Se diseñó una interfaz web usando php para extraer los resultados del modelo de minería aplicando las técnicas documentadas en la fase de modelado, a fin de recrear un simulador del proceso.

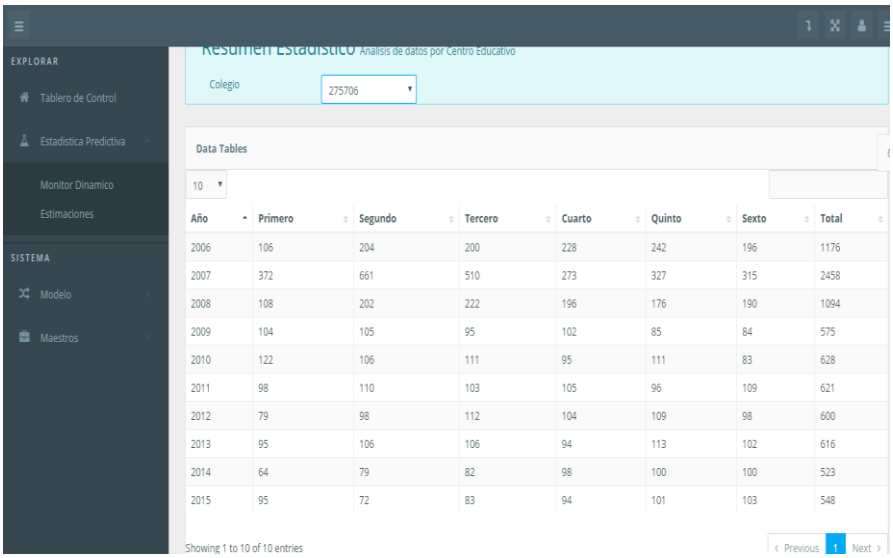
Interfaces relevantes del sistema

A. Interfaz de Logueo a simulador

B. Análisis de un determinado Colegio.

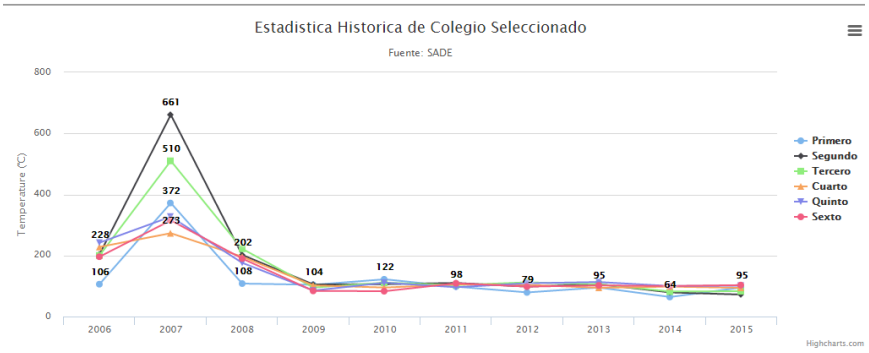


C. Análisis de Descomposición vista R Project a interfaz PHP



Año	Primero	Segundo	Tercero	Cuarto	Quinto	Sexto	Total
2006	106	204	200	228	242	196	1176
2007	372	661	510	273	327	315	2458
2008	108	202	222	196	176	190	1094
2009	104	105	95	102	85	84	575
2010	122	106	111	95	111	83	628
2011	98	110	103	105	96	109	621
2012	79	98	112	104	109	98	600
2013	95	106	106	94	113	102	616
2014	64	79	82	98	100	100	523
2015	95	72	83	94	101	103	548

D. Generación de gráficos de la representación matricial



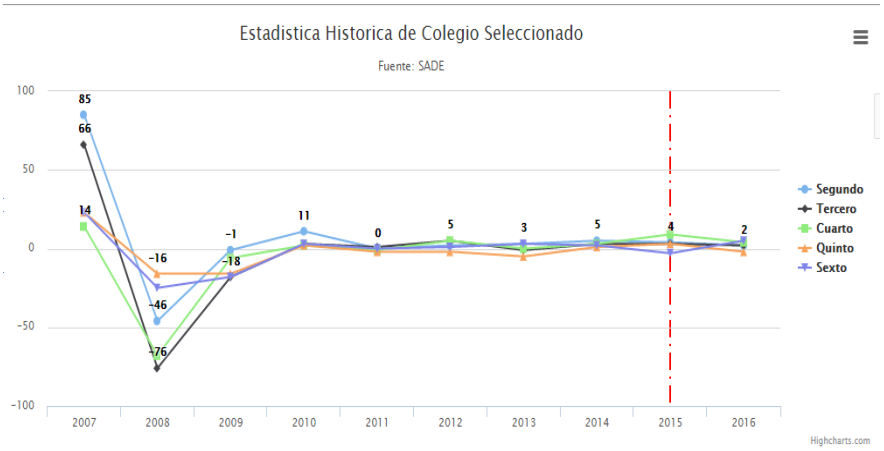
E. Pronósticos

Visualización General de los resultados

Año	Segundo	Tercero	Cuarto	Quinto	Sexto
2007	85	66	14	23	23
2008	-46	-76	-68	-16	-25
2009	-1	-18	-6	-16	-18
2010	11	3	2	2	3
2011	0	1	-1	-2	0
2012	2	5	5	-2	1
2013	3	-1	0	-5	3
2014	5	3	3	1	2
2015	4	3	9	3	-3
2016	2	2	4	-2	5

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Visualización por cada colegio



CAPITULO VI
CONCLUSIONES Y RECOMENDACIONES

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

a. Se recopilo y analizo la información brindada por la Ugel Chiclayo, se hizo el análisis de los archivos ofimáticos con los datos históricos de todos los alumnos matriculados, donde se determinó que se tuvo que homogenizar los campos, donde inicialmente comprendía 4831254 registros obtenidos están entre los años 2006 hasta 2015.

b. Se realizó la selección de las técnicas predictivas de minería de datos, determinando que el modelo a utilizar sería uno de series de tiempo y redes neuronales, dada la naturaleza de los datos analizados en el datawarehouse, se realizó un breve análisis de las técnicas u algoritmos que intervenían en este tipo de modelo como se muestra en la tabla 13.

c. Dentro de las técnicas predictivas se determinó utilizar los algoritmos de Redes Neuronales y ETS, ya que al realizar el análisis como se muestra en la tabla 14 se descartaron algunas técnicas adicionales por no tener los criterios necesarios para su implementación en el modelo a desarrollar.

d. Se realizó el análisis comparativo de técnicas de minería de datos con lo cual se demostró que para esta investigación las de series temporales se ajusta a nuestro estudio para lo cual dicha comparación y análisis se muestra en la tabla N° 14, de acuerdo a los criterios de selección se obtuvo que para el presente trabajo de investigación las técnicas más adecuadas son ETS y redes neuronales. Siendo las redes neuronales autoregresiva el que mejor confiabilidad presenta, Tanto para el nivel primario y secundario con

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

un 91% y 96% respectivamente.Podemos decir que Red neuronal autoregresiva obtuvo el nivel de confianza más elevado en comparación a ETS.

En el tiempo de procesamiento al evaluar estas técnicas se obtuvo que con el método red neuronal autoregresiva el tiempo promedio de ejecución de 3.355 segundos siendo superior a diferencia de ETS, que tiene 5,08.

e. Se construyó una aplicación web para evaluar los resultados obtenidos. El sistema se diseñó en php obteniendo una interfaz donde es capaz de interactuar con el servidor a fin de ejecutar los modelos ya sean reales o simulaciones donde se extrae el histórico de cada colegio analizado del datawarehouse,para que el usuario realice las pruebas pertinentes.

6.2. Recomendaciones

a. Se recomienda un licenciamiento en cuanto para el software SPSS.Para el tratamiento de los datos.
b. Los tratamientos de valores nulos en datos de esta naturaleza deben ser tratados con el mayor detalle posible, una matriz consolidada permitió identificar los valores faltantes en la base de datos que podían ocasionar daños en los cálculos de la serie.
c. Se recomienda que los formatos ofimáticos deberían de estar en un formato homogéneo.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

BIBLIOGRAFÍA

López Puga, J. (2010). INTRODUCCIÓN AL ANÁLISIS DE DATOS CON R Y R COMMANDER EN PSICOLOGÍA Y EDUCACIÓN. Bogotá, Colombia.

Bachman, J., Green, S., & Wirtanen, I. (1971). *Dropping out: Problem or symptom? Ann Arbor.* Michigan: Institute for Social Research, University of Michigan.

Barrientos, F., & Ríos, S. (2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. 1-36.

Brachman, R., & Anand, T. (1996). The process of Knowledge Discovery in Databases: A human centered approach. Advances in Knowledge Discovery and Data Mining. *AAAII MIT Press*.

Cabena, P. H. (1998). *Discovering Data Minin:From Concepts to Implementation.* New Jersey: Prentice Hall Saddle River.

Carrasco, R. A. (2011). *Data Mining: Aplicaciones Económico-Financieras.* España: Académica Española.

Cruz Arrela, L. (2010). *Minería de datos con aplicaciones.* Mexico: Universidad Nacional Autonoma de Mexico.

El Comercio. (26 de 06 de 2013). *Más de 8.000 escolares abandonaron las aulas durante el 2013*.

Elias, R., & Molina, J. (2005). *La deserción escolar de adolescentes en Paraguay.* Asuncion,Paraguay: Instituto de Desarrollo.

Española, R. A. (s.f.). *Diccionario de la Real Academia Española.*

Espíndola, E., & León, A. (2002). Éducación y conocimiento: una nueva mirada. *OEI*, 62.

Esteve, Juan Domingo. (s.f.). *http://platea.pntic.mec.es/vgonzale/cyr_0708/archivos/_15/Tema_5.6.htm.* Obtenido de http://platea.pntic.mec.es/vgonzale/cyr_0708/archivos/_15/Tema_5.6.htm

Fitzpatrick, K., & Yoels, W. (1992). Policy, school structure, and sociodemographic effects on statewide high school dropout rates. En K. Fitzpatrick, & W. Yoels, *Policy, school structure, and sociodemographic effects on statewide high school dropout rates* (págs. 76-93). Alabama: US: American Sociological Assn.

Flores, H. D. (2009). *Detección de Patrones de Daños y Averías en la Industria.* Buenos Aires.

Formia, S., Lanzarini, L., & Hasperué, W. (2013). Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos.

George Lee ;H. Jacky Chang,. (s.f.). *http://www.solociencia.com/ingenieria/07071201.htm.* Obtenido de http://www.solociencia.com/ingenieria/07071201.htm

Getoor, L., & Ben, T. (2007). *Introducción a estadística de relación de aprendizaje.* MIT.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Grudnitsky, B. J. (1992). *Diseño de sistemas de información. Teoría y Práctica.* México: Megabyte Grupo Noriega.

Guallart Romeu, P. M. (2010). MINERÍA DE DATOS APLICADA AL ANÁLISIS DEL TRATAMIENTO INFORMATIVO DE LA DROGADICCIÓN. 28.

Hand, M. &. (2011). *Principles of Data Mining.* Cambridge: MIT Press Cambridge.

Hernández & Ferri, C. (2004). *Introducción a la Minería de Datos.* España: Pearson.

Hyndman, R. (11 de Enero de 2015). *github.com*. Obtenido de github.com: https://github.com/robjhyndman/forecast

Hyndman, R. J. (2014). *Forecasting: principles and practice.* OTexts.

Irizarry, R., & Quintero, A. (2006). *ESTUDIOS DE CASOS NACIONALES: PUERTO RICO*.

Joshi, K. (1997). Analysis of data mining algorithms. *University of Minnesota*.

Lavado, P., & Gallegos, J. (2005). *La dinámica de la deserción escolar en el Perú:un enfoque usando modelos de duración.* Lima: Grade.

Lopez Alfonso, J. (06 de 02 de 2015). *Redes Neuronales*. Obtenido de Lopez Alfonso, Jesus: http://members.tripod.com/jesus_alfonso_lopez/RnaIntro2.html

Márquez, C., Romero, C., & Ventura, S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. 1.

Mazo, C. X., & Bedoya, O. (2010). PESPAD: una nueva herramienta para la predicción de la estructura secundaria de la proteína basada en árboles de decisión. *Ingeniería y Competitividad*, 9-22.

Ministerio de Educación, el 14% de niños y jóvenes entre los 13 y 19 años dejó el colegio o nunca se matriculó. (18 de Junio de 2014). *ProExpansion*.

Moody , J., & Darken, C. (1989). Fast Learning in networks of locally tuned processing .

Mora, S. L. (2002). *Programación de aplicaciones web.* Alicante, Argentina: Editorial Club Universitario.

Mora, S. L. (2002). *Programación de aplicaciones web .* Alicante, Argentina: Facultad de Ingeniería - Universidad de Buenos Aires.

Morrow, G. (1985). *Standardizing Practice in the Analysis of School Dropouts.* Columbia: Teachers College, Columbia University.

Msc. Marvin Lemos, AJ Alves, Douglas S. Kridi, Kannya Leal . (2015). *https://github.com/zerokol/.* Obtenido de https://github.com/zerokol/eFLL

Olabe Basogain, X. (2008). *Redes Neuronales Artificiales y sus apliciones.* España: Escuela Superior de Ingenieria de Bilbao.

oocities. (s.f.). *Sistemas y herramientas de minería de datos*. Recuperado el 22 de 08 de 2014, de http://www.oocities.org/es/mineria.datos/sistemas_herramientas_mineria_datos.pdf

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Orallo Hérnandez, J. (12 de 12 de 2015). *Minería de Datos.* Obtenido de http://users.dsic.upv.es/~jorallo/master/dm5.pdf

Ortiz Farro, P. (2015). *Minería de datos con series de tiempo en el desarrollo e implementacíon del sistema inteligente que predice la produccíon de arroz en el ámbito de la gerencia regional de Agricultura.* Chiclayo.

Pérez, C., & Santín, D. (2008). *Minería de Datos:Técnicas y Herramientas.* España: Thompson Ediciones Paraninfo,S.A.

Pernía, A., & F., C. (2001). Gestión del Conocimiento y Minería de datos. *XVII Congreso Nacional de Ingeniería de Proyectos, Murcia, Anonymous* .

Quinlan, J. R. (1993). *C4. 5: programs for machine learning.* Morgan Kaufmann.

Silvaz Wanumen, L. (2010). Minería de datos para la predicción de fraudes en tarjetas de crédito. 1-14.

Sinnexus. (s.f.). *Minería de datos*. Recuperado el 25 de 08 de 2014, de http://www.sinnexus.com/business_intelligence/datamining.aspx

Spositto, O., Etcheverry, M. E., Ryckeboer, H., & Bossero, J. (2010). Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. 1-5.

Timarán, R., Calderón, A., & Jiménez, J. (2013). Aplicación de la minería de datos en la extraccion de perfiles de desercion estudiantil.

Unicef. (s.f.). Recuperado el 19 de Agosto de 2014, de http://www.unicef.org/peru/spanish/children_3787.htm

Universidad de Barcelona. (s.f.). *TÉCNICAS DE PREDICCIÓN*. Recuperado el 22 de 08 de 2014, de http://www.ub.edu/aplica_infor/spss/cap8-5.htm

Usama, A., & Wierse, G. (2002). *Information and Visualization in Data Mining and Knowledge Diskovery.* Morgan Kauffmann.

Valero, S., Salvador, A., & García, M. (2003). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. 1-8.

Vílchez García, V. (2010). *Estimación y clasificación de daños en materiales utilizando modelos AR y redes neuronales para la evaluaciónno destructiva con ultrasonidos.* Granada.

WebMining Consultores. (10 de 01 de 2014). *KDD: Proceso de Extracción de conocimiento*. Recuperado el 12 de 07 de 2014, de http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

Weiss, S., & Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide.* EE UU: Morgan Kaufmann.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ANEXO

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

ANEXO 01: Diccionario de datos

CÓDIGO MODULAR:

CAMPO	TIPO	LONG	VALOR
Cod_mod	C	7	7 dígitos
DESCRIPCION			
El código modular es un número aleatorio que obedece a un algoritmo matemático. Cumple la función de identificar al centro educativo. Consta de 7 digitos, autogenerado por el software SIE (Sistema de Información Estadística). No presenta valores en blanco ni nulos.			

NÚMERO DE ANEXO:

CAMPO	TIPO	LONG	VALOR
Anexo	C	1	(*)
DESCRIPCION			
Identifica si el C.E. es principal o es un anexo. No presenta valores en blanco ni nulos.			

(*) Número de anexo:

"0" : Centros Educativos Principales.
"1" : Primer anexo.
"2" : Segundo anexo.
"3" : Tercer anexo.
"4" : Cuarto anexo.
"5" : Quinto anexo.
"6" : Sexto anexo.
"7" : Séptimo anexo.
"8" : Octavo anexo.
"9" : Noveno anexo.

NÚMERO DE CÉDULA:

CAMPO	TIPO	LONG	VALOR
Nroced	C	2	(*)
DESCRIPCION			
Identifica el número de cédula. No presenta valores en blanco ni nulos.			

(*) Se especifica para cada cédula en el presente diccionario de datos.
Los valores son los siguientes:

"01" : Cédula 1 – Educación Básica Regular - Inicial Escolarizada.
"02" : Cédula 2 – Educación Básica Regular - Inicial No Escolarizada.
"03" : Cédula 3 – Educación Básica Regular Primaria o Secundaria.
"04" : Cédula 4 - Educación Primaria o Secundaria de Adultos.
"4A" : Cédula 4-A – Educación Básica Alternativa – CEBA.
"05" : Cédula 5 - Educación Superior Pedagógica y Artística.
"06 : Cédula 6 - Educación Superior Tecnológica.
"08" : Cédula 8 - Educación Especial Escolarizada y No Escolarizada.
"09" : Cédula 9 - Educación Técnico Productiva.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

FORMA DEL CENTRO O PROGRAMA EDUCATIVO:

CAMPO	TIPO	LONG	VALOR
Formas	C	1	(*)
DESCRIPCION			
Forma en que se imparte la enseñanza en el Centro o Programa Educativo. No presenta valores en blanco ni nulos.			

(*) Formas:

"S" : Escolarizado.
"N" : No Escolarizado.

CODIGO DE UGEL:

CAMPO	TIPO	LONG	VALOR
Codooii	C	6	(*)
DESCRIPCIÓN			
Representa al código de Regiones y Ugels. No presenta valores en blanco ni nulos.			

(*) Ver codificador de Regiones y Ugels 2005.

CÓDIGO DE ÁREA:

CAMPO	TIPO	LONG	VALOR
Cod_area	C	2	(*)
DESCRIPCIÓN			
Representa al código de área donde se encuentra ubicado el centro o programa educativo. No presenta valores en blanco ni nulos.			

"10" : Urbano.
"11" : Urbano residencial.
"12" : Urbano marginal.
"13" : Urbano AA.HH. o pueblo joven.
"14" : Rural.

NIVEL Y/O MODALIDAD:

CAMPO	TIPO	LONG	VALOR
Niv_mod	C	2	(*)
DESCRIPCIÓN			
Representa al nivel y/o modalidad del centro o programa educativo. No presenta valores en blanco ni nulos			

(*) Código de Nivel y/o modalidad:

"A1" : Educación Básica Regular – Inicial Cuna.
"A2" : Educación Básica Regular – Inicial Jardín.
"A3" : Educación Básica Regular – Inicial Cuna Jardín.
"B0" : Educación Básica Regular – Primaria de Menores.
"C0" : Educación Básica Regular – Primaria de Adultos.
"F0" : Educación Básica Regular – Secundaria de Menores.
"G0" : Educación Básica Regular – Secundaria de Adultos.
"D0" : Educación Básica Alternativa – CEBA.
"K0" : Formación Magisterial ISP.

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

NÚMERO DE CUADRO:

CAMPO	TIPO	LONG	VALOR
Cuadro	C	4	(*)
DESCRIPCIÓN			
Identifica el número de cuadro de cada cédula esta ajustado con ceros a la derecha. No presenta valores en blanco ni nulos.			

(*) Se especifica para cada cuadro en el presente diccionario de datos.

TIPO DE DATO:

CAMPO	TIPO	LONG	VALOR
Tipdato	C	(**)	(*)
DESCRIPCIÓN			
Variable que toma valores de acuerdo al tipo de cédula y cuadro correspondiente. Los valores se detallan para cada tabla. Representa a la descripción vertical de cada cuadro de cada cédula. No presenta valores en blanco ni nulos.			

Anexo 02: Extracción del Conocimiento

Adquisición de la data

COD_MC	ANEXO	NROCE	CUAI	TIPDATO	DATO00	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	DAT	FORM	CODGE	CODOC	CC	NI	GE	GES	STAT	COD_(	AREA	FTE_
1314145	0	01	2200	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	A2	B4	1	3	1	INEI-
1314145	0	01	2300	1	0	0	4	2	1	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	A2	B4	1	3	1	INEI-
1302009	0	03	2200	1	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	2	0	0	0	0	28	0	41	0	32	0	26	0	36	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	3	0	0	0	0	24	0	35	0	26	0	20	0	28	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	4	0	0	0	0	4	0	6	0	6	0	6	0	8	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	5	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2200	10	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2310	2	0	0	3	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2320	2	0	0	3	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
1302009	0	03	2500	1	0	0	28	0	30	0	42	0	32	0	26	0	36	0	0	0	0	0	0	0	0	S	010101	010001	10	B0	A1	1	3	1	INEI-
0262196	0	03	2200	1	0	0	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	F0	A1	1	3	1	INEI-
0262196	0	03	2200	2	0	0	0	0	159	0	99	0	115	0	99	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	F0	A1	1	3	1	INEI-
0262196	0	03	2200	3	0	0	0	0	137	0	90	0	106	0	95	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	F0	A1	1	3	1	INEI-
0262196	0	03	2200	4	0	0	0	0	22	0	9	0	9	0	4	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	F0	A1	1	3	1	INEI-
0262196	0	03	2200	5	0	0	0	0	4	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	F0	A1	1	3	1	INEI-
0262196	0	03	2200	6	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	S	010101	010001	10	F0	A1	1	3	1	INEI-

Como se muestra en la imagen es la data inicial, como la ugel de Chiclayo entrego para la presente investigación.

Selección de variable

COD_	NRO	CU	TIPDAT	DATO	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	DA	FO	COD	COD	NI	G	ST	CO	ARE	FTE_AI
0345280	03	2200	1	0	30	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	2	0	0	0	33	40	34	43	32	35	40	25	22	26	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	3	0	0	0	28	35	31	39	29	32	39	24	22	26	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	4	0	0	0	5	5	3	4	3	3	1	1	0	0	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	5	0	0	0	4	4	3	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	6	0	0	0	4	3	3	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	7	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	8	0	6	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0345280	03	2200	10	0	6	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	S	140113	140001	10	B0	A1	1	3	1	INEI-D
0456467	03	2200	1	0	42	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	2	0	0	0	47	38	61	58	56	58	58	63	44	54	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	3	0	0	0	45	38	55	55	53	58	54	62	43	53	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	4	0	0	0	2	0	6	3	3	0	4	1	1	1	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	5	0	9	10	4	12	8	8	10	12	8	6	0	1	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	6	0	0	0	4	11	8	8	10	12	8	6	0	1	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	7	0	9	10	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	8	0	10	8	7	4	4	4	3	7	7	1	4	1	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0456467	03	2200	10	0	10	8	7	4	4	4	3	7	7	1	4	1	0	0	0	0	0	0	0	0	S	140108	140001	10	B0	A1	2	3	1	INEI-D
0446286	03	2200	2	0	0	0	2	3	2	1	1	3	3	2	0	0	0	0	0	0	0	0	0	0	S	140109	140001	20	B0	A1	1	2	2	DRE
0446286	03	2200	3	0	0	0	1	3	2	1	1	2	3	2	0	0	0	0	0	0	0	0	0	0	S	140109	140001	20	B0	A1	1	2	2	DRE

En la imagen se muestra la selección de variables para trabajar las cuales fueron CODOOII=código de Ugel, para esto código de la Ugel 14001, NIV_MOD=modalidad, y en la modalidad es BO y FO, las cuales representa al nivel primario y secundario respectivamente, y también CUADRO=el número de cuadro que para el estudio se basó en el 2200

Tratamiento de datos Nulos

Año	CodigoColeg	COD_MOD	DATO01H	DATO01	DATO02	DATO02	DATO03	DATO03	DATO04	DATO04	DATO05	DATO05	DATO06	DATO06	Ref
2006	277041	0217026	6	3	18	10	14	14	22	6	22	6	12	12	0
2006	275706	0344820	51	55	86	118	116	84	82	146	92	150	86	110	0
2006	275711	0344838	81	89	146	136	126	132	140	118	168	150	110	120	0
2006	673673	0344846	71	62	140	132	154	216	186	192	194	170	190	228	0
2006	280109	0344853	47	39	110	132	132	92	122	114	108	134	84	100	0
2006	278074	0344879	27	29	76	40	68	46	100	66	74	102	48	78	0
2006	278578	0344887	62	85	166	150	176	190	192	194	198	210	142	188	0
2006	278597	0344903	40	35	84	66	96	78	72	72	66	70	50	62	0
2006	278601	0344911	88	76	208	242	202	230	220	198	236	224	352	204	0
2006	280034	0344929	25	21	92	82	56	56	66	64	92	92	72	90	0
2006	280114	0344945	22	27	68	82	86	92	84	88	80	72	66	46	0
2006	280128	0344960	8	3	22	20	32	16	14	10	26	14	28	14	0
2006	280500	0344978	31	29	90	88	52	100	120	78	92	96	94	72	0
2006	280519	0344986	42	40	118	96	76	104	94	130	94	104	64	82	0
2006	280604	0345009	52	52	106	82	124	108	130	114	86	98	106	90	0
2006	275495	0345017	196	0	506	0	554	0	614	0	700	0	766	0	6
2006	275725	0345025	101	97	252	224	204	332	258	272	166	326	156	368	0
2006	275532	0345033	23	19	38	26	60	36	26	22	46	38	40	36	0
2006	275730	0345041	38	41	128	116	108	134	116	126	96	140	150	104	0
2006	275749	0345066	35	38	88	76	70	98	80	88	56	100	76	86	0
2006	279436	0345082	25	22	48	44	34	40	62	76	72	58	72	56	0
2006	279422	0345108	2	2	0	16	2	4	12	8	20	12	12	8	1
2006	279403	0345116	3	5	8	14	12	8	8	6	2	18	18	6	0
2006	279568	0345132	13	6	14	20	16	18	18	22	22	18	18	14	0
2006	279573	0345140	52	37	58	78	48	60	64	38	48	58	36	54	0
2006	279587	0345157	15	10	16	14	20	28	18	22	32	28	14	30	0

Para el tratamiento de los datos nulos se utilizó la formula =+CONTAR.BLANCO(D3:O3)+CONTAR.SI(D3:O3;0), donde solo se tomarían los que tengan como valor 0.

Pasando Datos Al Spss

Tablas personalizadas

NIV_MOD B0

		DAT001H	DAT001M	DAT002H	DAT002M	DAT003H	DAT003M	DAT004H	DAT004M	DAT005H	DAT005M	DAT006H	DAT006M
		Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma	Suma
COD_MOD	0217026	12	7	35	15	21	21	33	9	33	9	18	18
	0344820	129	132	335	183	174	126	123	220	138	225	129	165
	0344838	162	178	398	207	189	198	210	177	252	225	165	180
	0344846	142	124	339	202	231	324	279	288	291	255	285	342
	0344853	94	78	253	229	201	145	184	171	162	201	126	150
	0344879	49	63	168	62	102	69	150	99	111	153	72	117
	0344887	118	168	497	239	264	285	288	291	297	315	213	282
	0344895	38	48	146	97	93	118	115	130	105	93	123	117
	0344903	80	70	182	110	152	119	108	108	100	106	75	93
	0344911	175	196	573	363	303	345	330	297	354	336	528	306
	0344929	61	116	162	123	84	84	99	96	138	138	108	135
	0344945	44	57	165	129	133	140	126	133	120	108	99	69
	0344960	28	18	55	30	48	24	21	15	39	21	42	21
	0344978	62	58	193	135	81	151	181	117	138	144	141	108
	0344986	97	97	260	164	117	158	143	196	141	156	96	123
	0345009	102	98	273	139	186	162	195	171	129	147	159	135
	0345017	202	0	1151	0	921	0	921	0	1059	0	1140	0

Luego de haber tratado los datos nulos se pasó toda la data a la herramienta spss para hacer un pivoteo.

Migrando Del Spss al excel

Año	CodigoColegio	COD_MOD	DATO01H	DATO01M	DATO02H	DATO02M	DATO03H	DATO03M	DATO04H	DATO04M	DATO05H	DATO05M	DATO06H	DATO06M	Nivel
2006	277041	0217026	6	3	18	10	14	14	22	6	22	6	12	12	BO
2006	275706	0344820	51	55	86	118	116	84	82	146	92	150	86	110	BO
2006	275711	0344838	81	89	146	136	126	132	140	118	168	150	110	120	BO
2006	673673	0344846	71	62	140	132	154	216	186	192	194	170	190	228	BO
2006	280109	0344853	47	39	110	132	132	92	122	114	108	134	84	100	BO
2006	278074	0344879	27	29	76	40	68	46	100	66	74	102	48	78	BO
2006	278578	0344887	62	85	166	150	176	190	192	194	198	210	142	188	BO
2006	278597	0344903	40	35	84	66	96	78	72	72	66	70	50	62	BO
2006	278601	0344911	88	76	208	242	202	230	220	198	236	224	352	204	BO
2006	280034	0344929	25	21	92	82	56	56	66	64	92	92	72	90	BO
2006	280114	0344945	22	27	68	82	86	92	84	88	80	72	66	46	BO
2006	280128	0344960	8	3	22	20	32	16	14	10	26	14	28	14	BO
2006	280500	0344978	31	29	90	88	52	100	120	78	92	96	94	72	BO
2006	280519	0344986	42	40	118	96	76	104	94	130	94	104	64	82	BO
2006	280604	0345009	52	52	106	82	124	108	130	114	86	98	106	90	BO
2006	275725	0345025	101	97	252	224	204	332	258	272	166	326	156	368	BO
2006	275532	0345033	23	19	38	26	60	36	26	22	46	38	40	36	BO
2006	275730	0345041	38	41	128	116	108	134	116	126	96	140	150	104	BO
2006	275749	0345066	35	38	88	76	70	98	80	88	56	100	76	86	BO
2006	279436	0345082	25	22	48	44	34	40	62	76	72	58	72	56	BO

Una vez realizado el pivoteo en la herramienta Spss se pasó a un formato homogéneo en Excel donde estarán alojados todos los registros por año así como se muestra en la imagen

Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

Migrando Del Excel a Sql Server

	Año	CodigoColegio	COD_MOD	DATO01H	DATO01M	DATO02H	DATO02M	DATO03H	DATO03M	DATO04H	DATO04M	DATO05H	DATO05M	DATO06H	DATO06M	Nivel
1	2006	277041	0217026	6	3	18	10	14	14	22	6	22	6	12	12	BO
2	2006	275706	0344820	51	55	86	118	116	84	82	146	92	150	86	110	BO
3	2006	275711	0344838	81	89	146	136	126	132	140	118	168	150	110	120	BO
4	2006	673673	0344846	71	62	140	132	154	216	186	192	194	170	190	228	BO
5	2006	280109	0344853	47	39	110	132	132	92	122	114	108	134	84	100	BO
6	2006	278074	0344879	27	29	76	40	68	46	100	66	74	102	48	78	BO
7	2006	278578	0344887	62	85	166	150	176	190	192	194	198	210	142	188	BO
8	2006	278597	0344903	40	35	84	66	96	78	72	72	66	70	50	62	BO
9	2006	278601	0344911	88	76	208	242	202	230	220	198	236	224	352	204	BO
10	2006	280034	0344929	25	21	92	82	56	56	66	64	92	92	72	90	BO
11	2006	280114	0344945	22	27	68	82	86	92	84	88	80	72	66	46	BO
12	2006	280128	0344960	8	3	22	20	32	16	14	10	26	14	28	14	BO
13	2006	280500	0344978	31	29	90	88	52	100	120	78	92	96	94	72	BO
14	2006	280519	0344986	42	40	118	96	76	104	94	130	94	104	64	82	BO
15	2006	280604	0345009	52	52	106	82	124	108	130	114	86	98	106	90	BO
16	2006	275725	0345025	101	97	252	224	204	332	258	272	166	326	156	368	BO
17	2006	275532	0345033	23	19	38	26	60	36	26	22	46	38	40	36	BO
18	2006	275730	0345041	38	41	128	116	108	134	116	126	96	140	150	104	BO
19	2006	275749	0345066	35	38	88	76	70	98	80	88	56	100	76	86	BO
20	2006	279436	0345082	25	22	48	44	34	40	62	76	72	58	72	56	BO
21	2006	279422	0345108	2	2	0	16	2	4	12	8	20	12	12	8	BO
22	2006	279403	0345116	3	5	8	14	12	8	8	6	2	18	18	6	BO
23	2006	279568	0345132	13	6	14	20	16	18	18	22	22	18	18	14	BO
24	2006	279573	0345140	52	37	58	78	48	60	64	38	48	58	36	54	BO
25	2006	279587	0345157	15	10	16	14	20	28	18	22	32	28	14	30	BO

Una vez realizado la migración del spss al Excel, se procedio a migrar la data al sql para realizar más adelante la aplicación de los algoritmos seleccionaos anteriormente.

Anexo 03: Laboratorio

Evaluación de tiempo de Holwinter

```
Loading required package: timeDate
This is forecast 6.2

>
> z<- c(15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,8$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
> consumo<-ts(z, start=c(2011,3),frequency=12)
> seasonplot(consumo)
>
>
>
> HW<-HoltWinters(consumo)
>
> system.time(HoltWinters(consumo))
   user  system elapsed
   0.02    0.00    0.02
>
```

Evaluación de tiempo de Red Neuronal Autoregresiva

```
> z<- c(15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,8$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
+ 15,7,7,11,14,16,15,13,12,11,9,11,9,9,8,0,4,26,24,51,42,41,77,69,69,77,83,75,7$
> consumo<-ts(z, start=c(2011,3),frequency=12)
> seasonplot(consumo)
>
> nnetar(z,2,P=1,2,repeats=20)
Series: z
Model:  NNAR(2,2)
Call:   nnetar(x = z, p = 2, P = 1, size = 2, repeats = 20)

Average of 20 networks, each of which is
a 2-2-1 network with 9 weights
options were - linear output units

sigma^2 estimated as 148
> system.time(nnetar(z,3,P=1,2,repeats=20))
   user  system elapsed
   0.32    0.00    0.33
```

Anexo 04: Evaluación Económica

Para hacer el cálculo del costo del software se utilizó el modelo COCOMO (**CO**nstructive **CO**st **MO**del)

ANÁLISIS PRELIMINAR

DEFINICIÓN DE REQUERIMIENTOS:

Donde:

RS = Responsabilidades del Sistema

Se considera la siguiente lista, siendo seis:

a. Generar modelo de series de tiempo
b. Entrenar modelo
c. Realizar estimaciones
d. Generar reportes
e. Visualizar comparación de modelos predictivos

$$\mathbf{RS = 5}$$

F = Funciones de Sistema:

$$F = 280 * RS$$

$$\mathbf{F = 1400}$$

MF = Miles de Funciones

$$\mathrm{MF} = \frac{\mathrm{F}}{1000}$$

$$\mathrm{MF} = \frac{1400}{1000}$$

$$\mathbf{MF = 1.4}$$

ESF = Esfuerzo.

$$\mathrm{ESF} = 2.4(\mathrm{MF})^{1.05}$$

$$\mathrm{ESF} = 2.4(1.4)^{1.05}$$

$$\mathbf{ESF = 3.97560561}$$

TDES = Tiempo de Desarrollo

$$\mathrm{TDES} = 2.5(\mathrm{ESF})^{0.38}$$

$$\mathrm{TDES} = 2.5(3.97560561)^{0.38}$$

$$\mathbf{TDES = 4.22}$$

CH = Cantidad de Hombres por MES

$$\mathrm{CH} = \mathrm{ESF}/\mathrm{TDES}$$

$$\mathrm{CH} = \frac{\mathbf{3.97560561}}{4.22}$$

$$\mathrm{CH} = 0.94208$$

$$\mathbf{CH = 1\ personas\ por\ mes}$$

CHM = Costo Hombre por Mes

$$\mathrm{CHM} = \mathrm{CH} * \mathrm{SPM\ (Salario\ Promedio\ Mensual)}$$

$$\mathrm{CHM} = 1 * 2400$$

$$\mathbf{CHM = 2400}$$

CD = Costo de Desarrollo

$$\mathrm{CD} = \mathrm{ESF} * \mathrm{CHM}$$

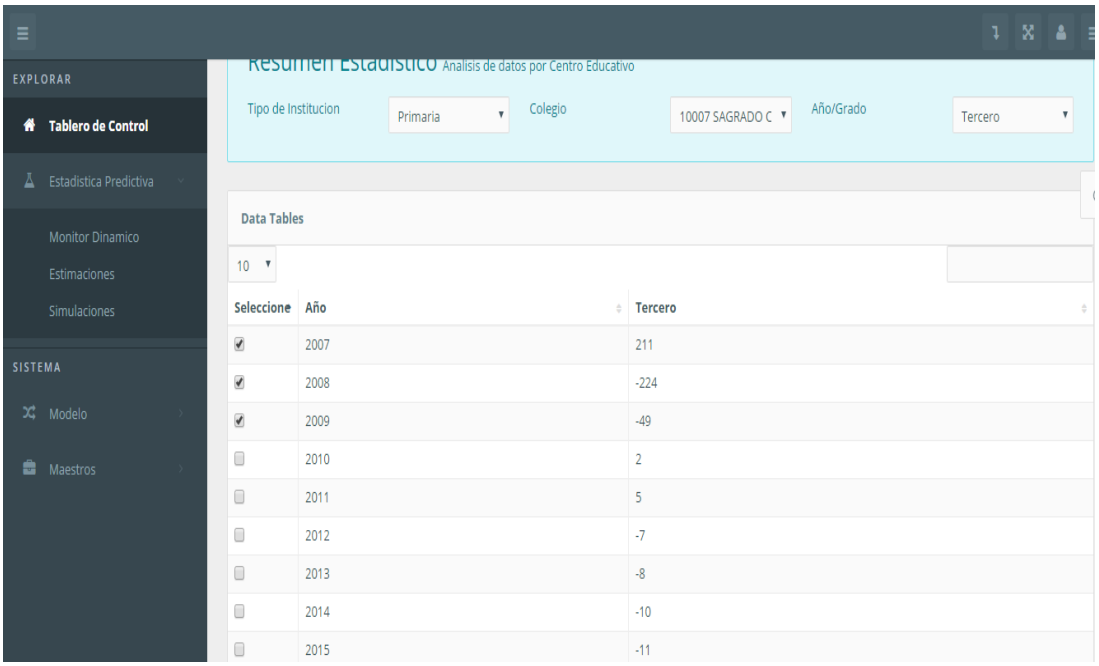
Campus Universitario Km. 5 Carretera a Pimentel - Chiclayo - Perú
Teléf: (+51)(74) 481610 / 481620 - Fax: 203861
www.uss.edu.pe

$$CD = 3.975 * 2400$$

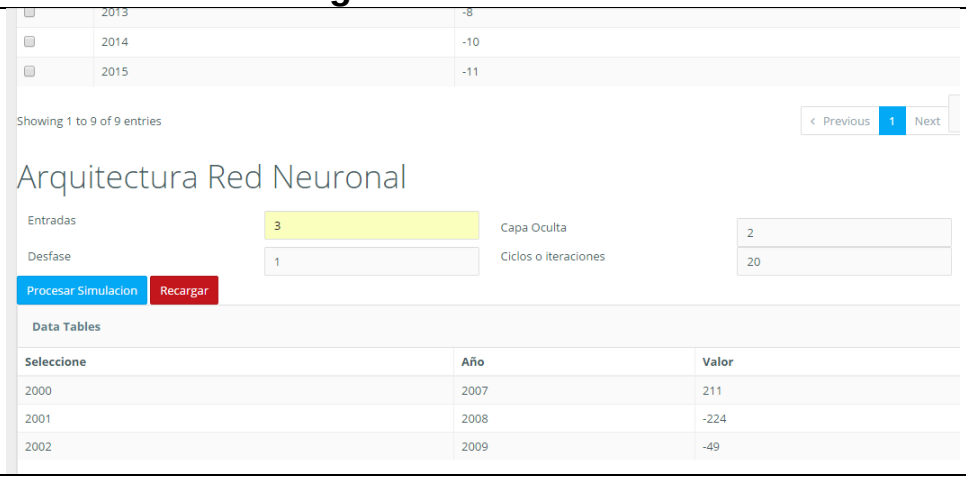
$$\mathbf{CD = S/.\ 9{,}480.00}$$

Los costos obtenidos serán asumidos en su totalidad por el responsable de la investigación.

Anexo 05: Simulación

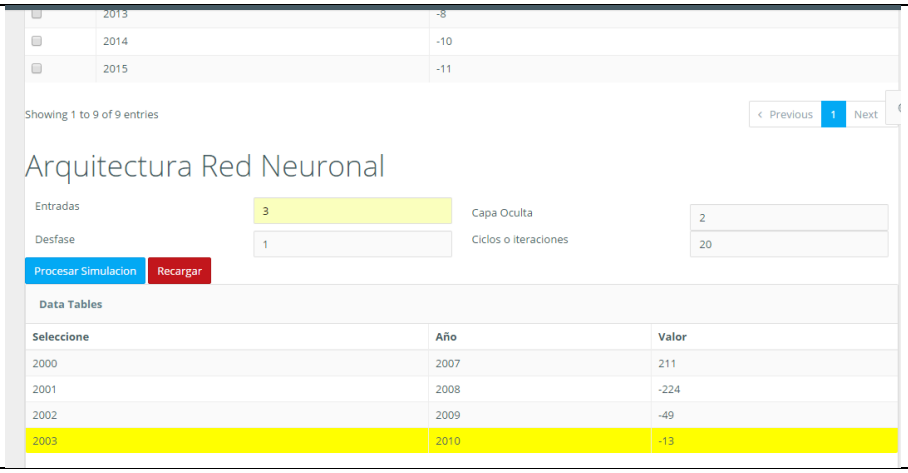
Módulo de Simulación

Como se muestra en la imagen se puede visualizar el módulo de simulación donde se tiene que seleccionar el nivel (Primaria o Secundaria), el colegio y el grado que se quiere simular. Donde también se selecciona los años que se quiere simular en este caso se va a simular tres años para predecir el año 2010. Donde en la Columna Tercero es La cantidad de Alumnos que han desertado donde luego se pasara a la red neuronal

Configurando Red Neuronal



En la imagen se muestra la configuración de la red neuronal, donde se ingresa la entrada en este caso se ingresan 3 entradas a la red neuronal, la capa oculta es 2 ,el desfase es 1 y el ciclo de repeticiones(Iteraciones). Luego se da clic en el botón procesar simulación y luego se da clic en el botón recargar

Resultados de Simulación



En la imagen se muestra la simulación que se hizo con el algoritmo.