UNIVERSIDAD PERUANA UNION

ESCUELA DE POSGRADO

MAESTRIA EN INGENIERIA DE SISTEMAS

Una Institución Adventista

ANÁLISIS PREDICTIVO BASADO EN REDES NEURONALES NO SUPERVISADAS APLICANDO ALGORITMO DE K-MEDIAS Y CRISP-DM PARA PRONÓSTICO DE RIESGO DE MOROSIDAD DE LOS ALUMNOS EN LA UNIVERSIDAD PERUANA UNIÓN

Tesis presentada para optar el Grado Académico de magíster en Ingeniería de Sistemas con mención Dirección y Gestión de Tecnologías de la Información

Por

Rodolfo Pacco Palomino

Lima, Perú, 2015

Análisis predictivo basado en redes neuronales no supervisadas aplicando algoritmo de K-MEDIAS y Crisp-DM para pronóstico de riesgo de morosidad de los alumnos de la Universidad Peruana Unión

TESIS

Presentada para optar el Grado Académico de Magíster en Ingeniería de Sistemas con mención en Dirección y Gestión de Tecnologías de Información

JURADO DE SUSTENTACIÓN

Dra. Erika Inés Acuña Salinas
Presidenta

Dr. Juan Jesús Soria Quijaite
Secretario

Dr. Guillermo Mamani Apaza
Asesor

Mg. Abel Zárate Avendaño
Vocal

Mg. Jorge Alejandro Sánchez Garcés
Vocal

Villa Unión, Ñaña, 05 de noviembre de 2015

DEDICATORIA

A mis padres, por su amor incondicional, para hacer realidad este proyecto en mi vida.

A mi hermano Abdón, quien es la persona indispensable para ser cada día mejor.

AGRADECIMIENTOS

Gracias a Dios todopoderoso,

quien han hecho posible este proyecto.

A la

Universidad Peruana Unión, por la oportunidad

que he tenido de estudiar allí y lograr

un desarrollo integral

A mi asesor,

Dr. Guillermo Mamani, por su orientación y apoyo

durante el desarrollo del proyecto.

A mis amigos(as),

por animarme a seguir adelante.

ÍNDICE GENERAL

LISTA DE FIGURAS

LISTA DE TABLAS

ANEXOS

LISTA DE ABREVIATURAS, SIGLAS Y ACRÓNIMOS

BA.	Business Analytics.
DSS.	Decision Support System.
TI.	Tecnologías e Información.
DW.	Datawarehouse
ANALYSIS SERVICE.	Herramienta BI de Microsoft.
CRISP-DM.	Cross Industry Standard Process for Data Mining.
ETL.	Extract, Transformation, and Load.
SQL SERVER.	Structured Query Language.
UPeU.	Universidad Peruana Unión.

RESUMEN

El presente trabajo de investigación de tesis desarrolla los indicadores de gerencia financiera, capturados de la necesidad de los clientes, éstos son modelados y desarrollados a través de las tecnologías de BA (Business Analytics), las cuales tienen el objetivo de mostrar los riesgos de morosidad. Este proyecto de investigación se ha desarrollado basado sobre redes neuronales y la metodología CRISP-DM, para implementar e implantar el proyecto de BA (Business Analytics). Se ha hecho una optimización del ciclo de vida de la metodología de CRISP-DM, según sus fases conocidas: comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación y despliegue.

El caso de estudio es el riesgo de morosidad de los alumnos de la Universidad Peruana Unión (UPeU), formado por cinco facultades: Ingeniería y Arquitectura, Ciencias de la Salud, Ciencias Empresariales, Ciencias Humanas y Educación y Teología. Para este estudio, el principal responsable del negocio es la Universidad Peruana Unión (UPeU).

En este proyecto de investigación de tesis se decide la herramienta de BI de Microsoft para el desarrollo de la solución y se elige la herramienta de Analysis Services. Como la solución de inteligencia de negocios se diseña los modelos de clúster, para la toma de decisión, utilizando las herramientas integration services para realizar ETL (Extraction Transform and Load).

En esta investigación se explica ampliamente que la implementación de un proyecto, utilizando la herramienta analysis services, consiste diferentes etapas de BI, desde el análisis de datos hasta los reportes de modelos de clasificación. Este proyecto servirá como base para elaborar proyectos de esta naturaleza o similares.

Palabras claves: riesgos de morosidad, redes neuronales y la metodología CRISP-DM.

ABSTRACT

This research thesis develops indicators Financial Management, captured the need of customers, these are shaped and developed through technologies BI (Business Intelligence). Which aim to show risk of dropping out. This research project has been developed based on neural networks and the CRISP-DM methodology to implement and deploy project BI (Business Intelligence). It has made a life cycle optimization methodology CRISP-DM phases as known as: Business understanding, data understanding, data preparation, modeling, evaluation and deployment.

The case study is the risk of students dropping out of (UPeU) (Peruvian Union University) which is made up with different powers such as: Engineering and Architecture, Health Sciences, Business, Education and Human Sciences finally Theology. The primary responsibility of business is the Peruvian Union University (UPeU).

In this thesis research project is decided by the tool for Microsoft BI solution development tool and select Analysis Services. As the Business Intelligence solution cluster models for decision making, is designed using the Integration Services tools for ETL (Extraction Transform and Load).

In this thesis research project is fully explained to the implementation of an Analysis Services project using the tool, implementation is at different stages of BI, from analyzing data to reports classification models. This project will serve as a basis for projects of this type or similar.

Key words: risks of delinquency, networks neuronales and the methodology CRISP-DM.

GLOSARIO

BUSINESS ANALYTICS. Se encarga de generar conocimientos sobre la base de una información, para esto se han creado herramientas de tecnologías que permiten construir soluciones de inteligencia de negocios.

EXECUTIVE INFORMATION SYSTEM. Provee a los gerentes de un acceso sencillo a la información interna y externa de su compañía.

CRISP-DM. El estándar incluye un modelo y una guía, estructurados en **seis fases**, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores.

ANALYSIS SERVICES. Analysis services es un motor de datos analíticos en línea que se usa en soluciones para ayudar en la toma de decisiones y Business Intelligence (BI).

K-MEDIAS. Es un método de agrupamiento que tiene como objetivo la partición de un conjunto de datos. Es un método para utilizar en minería de datos.

INTEGRATION SERVICES. Proporciona la solución ideal para cualquier tipo de integración de datos, análisis de negocio o proyecto de datos grande.

CAPÍTULO I

INTRODUCCIÓN A LA INVESTIGACIÓN

En esta tesis se busca las razones que explican el problema de la morosidad de los alumnos de la Universidad Peruana Unión. Los administradores necesitan formas más eficientes de analizar la información para tomar decisiones, y consecuentemente sobrevivir a los cambios que se producen en su entorno, los sistemas actuales no soportan las exigencias del negocio para generar reportes que apoyen en la toma de decisión, por esto surge la necesidad de crear una tecnología utilizando la herramienta de BA, para generar modelos de segmentación por las características similares que apoyan a la toma de decisiones.

El presente trabajo tiene el objetivo general de desarrollar una herramienta BA que sirve de soporte para la toma de decisiones en los procesos financieros académicos de la (UPeU), para esto se han implementado los modelos de segmentación para la Gerencia Financiera de la Universidad Peruana Unión. Estos indicadores del negocio son determinados por la necesidad del cliente.

A continuación se propone la herramienta para el análisis de datos, se realiza en tres etapas: definición, diseño de la decisión para la toma de decisiones. Luego el diseño del modelo de clasificación. Para la explotación de la información de manera sencilla para los usuarios que no conocen análisis y administración del negocio, y los análisis de datos serán representados en reportes dinámicos, donde se podrá analizarlos de distintas perspectivas según su necesidad.

El capítulo I de la tesis presenta la introducción de toda la investigación. El capítulo II trata los conceptos relacionados con la tecnología de BI, los cuales son explicadas de forma sencilla y detallada. El capítulo III presenta el método de la investigación basados en la metodología CRISP-DM, sobre las tareas de BI en las fases del proyecto. En el capítulo IV se presenta toda la información del proceso de

construcción para el análisis de datos financieros de los alumnos de la Universidad Peruana Unión realizadas con la herramienta Analysis Services. El capítulo V se explica la validación y los resultados dinámicos solicitados por el gerencia financiera de la (UPeU).

1.1 Identificación del problema

Los alumnos de la Universidad Peruana Unión provienen de diferentes lugares. Llegan con el propósito de estudiar las carreras que existe en la universidad, la gerencia financiera no tiene el conocimiento de segmentación de alumnos por grupos de morosidad, al momento de matricularse, el alumno asume pagar sus pensiones de enseñanza, divididas en 4 partes o armadas. La condición o situación de cada alumno es imposible conocer por parte de finanzas ya que el sistema actual de finanzas no reporta de modo adecuado y completo. Hoy en, día la mayoría de los estudiantes se autosostienen o son de pocos recursos económicos, no cuentan con apoyo de sus padres; en algunos casos, los alumnos reciben ayuda puntual de sus padres; sin embargo, el alumno simplemente no paga su pensión de enseñanza por motivos que se desconoce.

Al finalizar de cada semestre, el sistema actual reporta un buen porcentaje de deudas de los alumnos; esto afecta a la universidad generando grandes pérdidas y oportunidades.

1.2 Formulación del problema

Tomando en cuenta los problemas de pago que afronta la Universidad Peruana Unión, ¿cómo el análisis predictivo, basado en redes neuronales no supervisada, aplicando el algoritmo de K-Medias y Crisp-DM, ayuda mucho en el pronóstico del riesgo de morosidad de los alumnos de la Universidad Peruana Unión, año 2015?

1.3 Objetivo general

Implementar el análisis predictivo, basado en redes neuronales no supervisadas, aplicando el algoritmo de K-MEDIAS y Crisp-DM, para ayudar mucho en el pronóstico de riesgo de morosidad de los alumnos de la Universidad Peruana Unión, año 2015.

1.3.1 Objetivos específicos

Para implementar el análisis predictivo la propuesta de mejora se deberá realizar:

a) Realizar un diagnóstico de la situación actual en la que se encuentra la oficina de finanzas alumnos de la Universidad Peruana Unión.
b) Utilizar la metodología CRISP-DM, para realizar mejores análisis según las fases establecidas.
c) Segmentar a los alumnos de la Universidad Peruana Unión de acuerdo con su conducta de pagos.

1.4 Hipótesis

1.4.1 Hipótesis general

Si se aplica el análisis predictivo, basado en redes neuronales no supervisada, aplicando el algoritmo de K-Medias y Crisp-DM, entonces se ayudará mucho en el pronóstico de riesgo de morosidad de los alumnos de la Universidad Peruana Unión.

1.4.2 Hipótesis Específicas

a) Si se aplica el proceso de segmentación, basado en redes neuronales no supervisadas, y aplicando el algoritmo de K-Medias y Crisp-DM, entonces se ayudará mucho en el pronóstico de riesgo de morosidad de los alumnos de la UPeU año 2015.
b) Si se aplica el proceso de verificación de datos, basado en redes neuronales no supervisadas, apoyando en el algoritmo de K-Medias y Crisp-DM, entonces se ayudará mucho en el pronóstico de riesgo de morosidad de los alumnos de la UPeU, 2015.

1.5 Población.

La población para este proyecto está determinada por los alumnos de las diferentes escuelas profesionales de la UPeU, los cuales están divididos por:

Tabla Nro. 1, población.

E.A.P	CANTIDAD
Ingeniería y Arquitectura	67
Ciencias de la salud	10
Ciencias humanas y educación	17
Ciencias Empresariales	27
Teología	9

TOTAL	**130**

CAPÍTULO II

MARCO TEÓRICO

2.1 Antecedentes de la investigación.

2.1.1 Ámbito internacional.

Según J.R. Santos [1], en su investigación con el título de tesis EL NUEVO ACUERDO DE CAPITAL DE BASILEA ESTIMACIÓN DE UN MODELO DE CALIFICACIÓN DE PEQUEÑAS Y MEDIANAS EMPRESAS PARA EVALUAR EL RIESGO DE CRÉDITO, para optar el grado de Doctor.

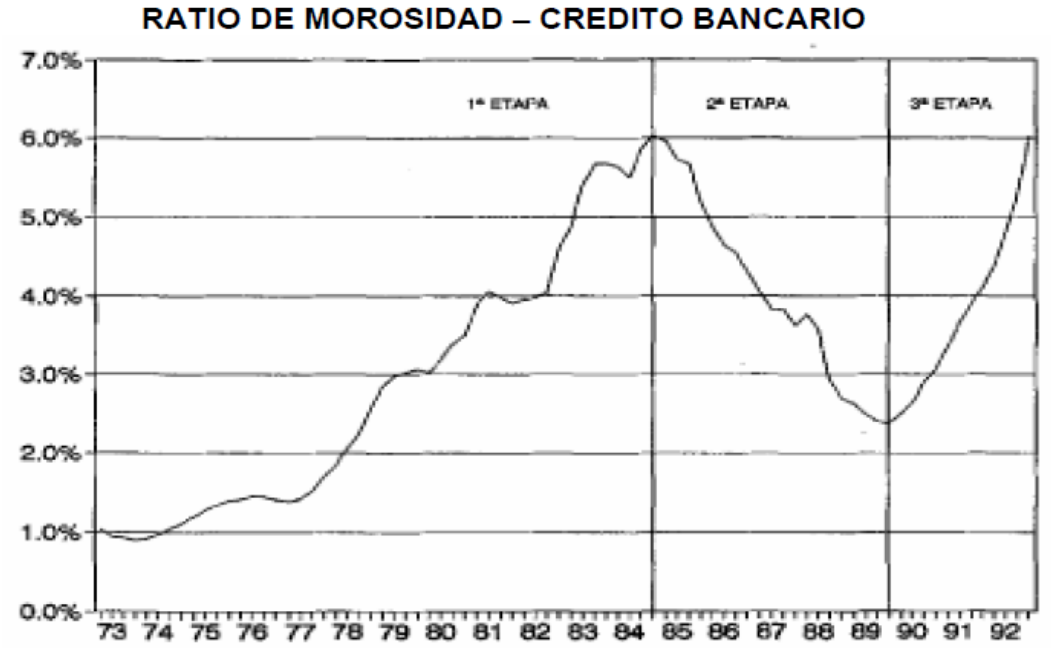
2.1.1.1 La morosidad Bancaria

El impacto del riesgo de crédito en las entidades se refleja, además en un aumento de la morosidad, en unas mayores necesidades de dotaciones a insolvencias. La estabilidad de una entidad o del conjunto del sistema bancario depende, en buena medida, de que dichas provisiones estén cubriendo adecuadamente la pérdida esperada de las carteras crediticias de las entidades. “La estabilidad del sistema financiero se ve influida por el riesgo empresarial en el sector real de la economía. El trabajo contiene un análisis exploratorio del riesgo empresarial y sus implicaciones para la morosidad del sector no financiero entre 1992 y 2002, a partir de las distribuciones empíricas de la rentabilidad contable y de la deuda bancaria morosa. Los resultados obtenidos muestran que, en cualquier fase del ciclo, un número significativo de empresas se encuentra en situación de riesgo económico y que la morosidad está efectivamente asociada con bajos beneficios de las empresas, sobre todo en las fases contractivas del ciclo económico. Por otra parte, el comportamiento de la morosidad agregada a lo largo del ciclo económico se explica mejor por

desplazamientos de deuda hacia empresas con dificultades financieras que por variaciones en la proporción de deuda morosa de las empresas con mora"[1].

Se ha de tener presente, en el análisis de la evolución de la morosidad, que la ratio es una resultante tanto del total de créditos morosos existentes en un determinado momento como del volumen total de créditos concedidos u otorgados.

La evolución trimestral de la ratio de morosidad, para el sector bancario, durante el período 1973-1992 puede verse en el Gráfico Nro. 1.



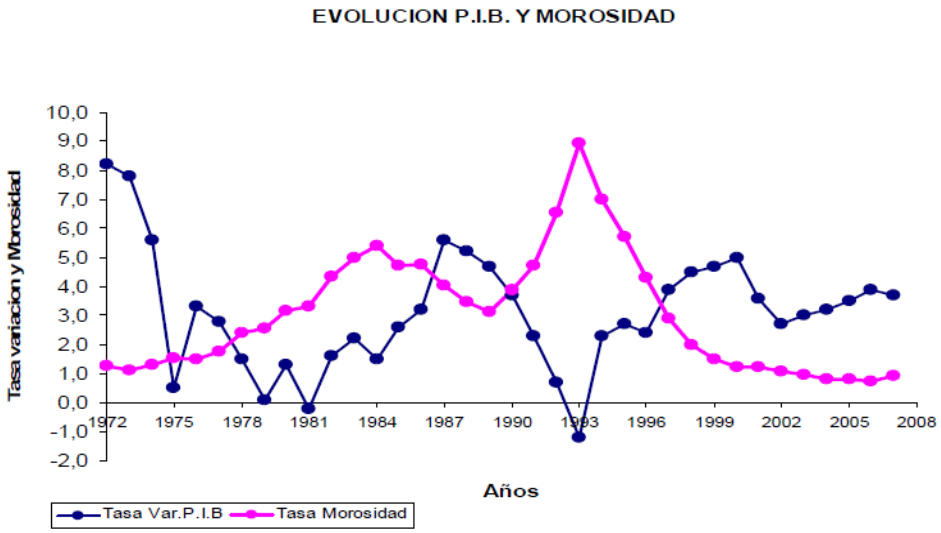
FUENTE: Freixas, X et al.[63]

Figura Nro. 1, Ratio de morosidad – crédito bancario.

En el gráfico anterior se pueden ver tres etapas claramente diferenciadas en la evolución tendencial de la ratio de morosidad: la primera se caracteriza por un crecimiento prácticamente continuo de la *ratio*, que pasa, de aproximadamente el 1% al principio de los años setenta, a algo más del 6% en el último trimestre de 1984 y el primero de 1985. No obstante, el proceso de crecimiento es mucho más lento al principio del período. En la segunda etapa, que cubre desde el segundo trimestre de 1985 al cuarto de 1989, se produce una continua caída de la ratio, que tiene su origen en la caída de la morosidad nominal. Sin embargo, no se alcanza niveles tan bajos como los del principio del período analizado [1].

Por último, la tercera etapa abarca el final de la muestra, se caracteriza igual que la primera, por el aumento de la *ratio,* aunque en esta ocasión dicho aumento es más rápido. La ratio de morosidad se define como el cociente entre activos dudosos e inversión crediticia, es el indicador tradicionalmente utilizado para analizar el riesgo de crédito en las entidades financieras y es asumida como medida fundamental de la calidad del crédito, por cuanto viene a reflejar el nivel de acierto de la entidad en la selección de sus clientes y en la solución de aquellos que devinieron en problemáticos [1].

La morosidad además de estar causada por la calidad con la que se gestiona el riesgo, puede estar causada por circunstancias económicas exógenas. Mientras en el primer caso las entidades crediticias pueden reducir su morosidad revisando sus políticas de riesgo y mejorando su gestión, poco pueden reducir la morosidad debido a factores exógenos a la propia entidad como el ciclo económico. Existe una relación estrecha entre la morosidad y el ciclo económico. Durante las fases recesivas crece la morosidad, consecuencia de las dificultades financieras de las empresas y de los hogares. Cuando la economía experimenta un fuerte crecimiento, los beneficios de las empresas no financieras y los ingresos de los hogares aumentan, lo que facilita el reembolso de los créditos, contribuyendo así al descenso de los ratios de morosidad de las entidades crediticias, pero en esta fase del ciclo suelen estar presentes políticas de crédito laxas y los riesgos no reciben el adecuado tratamiento. El Gráfico Nro. 2 ilustra sobre la correlación negativa existente entre la ratio de morosidad de las entidades de crédito y la tasa de crecimiento del PIB [1].



FUENTE: Contabilidad Nacional de España. INE
Boletín Estadístico del Banco de España
Elaboración propia

Figura Nro. 2, Evolución P.I.B. y Morosidad

Según el artículo "Software de minería de datos", el análisis de clúster es la determinación de la existencia de los grupos homogéneos. Los diferentes tipos de herramientas de clusterización que existe para el análisis de la información [2].

El estudio de "propagación de información en redes complejas de circuitos electrónicos no lineales con ruido", indica la cantidad de conexiones: los nodos que son conectadas tanto en la entrada y salida. La media de los índices de clúster de cada nodo es el índice de clúster de la red neuronal.

2.1.2 Ámbito nacional.

Elaboración de un mapeo de clústers en el Perú. Consultoría solicitada por el Consejo Nacional de la Competitividad [3].

2.1.2.1 Oportunidades del mercado

Las oportunidades son aquellos factores, positivos, que se generan en el entorno y que, una vez identificados, pueden ser aprovechados. Las principales oportunidades del mercado de congelados y conservas de pescado son:

a) Dado que el Perú tiene mayores y mejores relaciones con socios económicos extranjeros ahora tiene la posibilidad de atraer mayor inversión, transferencia de tecnología, intercambio de conocimientos y experiencias y la posibilidad de nuevos negocios [3].

b) Para el caso de conservas de pescado, la aparición de nuevos envases, ha contribuido a mejorar la imagen de estos productos. Además estos envases posibilitan productos más elaborados que pueden ofrecerse a precios más bajos.
c) La reciente tendencia hacia la promoción de la acuicultura dinamiza al clúster de congelados y conservas ya que, según la FAO, el 44% del pescado de consumo humano directo (del cual las conservas y congelados son parte importante) provienen de la acuicultura.
d) Existe una serie de proyectos viables bajo el marco del SNIP orientados a la remodelación, construcción o mejoramiento de desembarcaderos pesqueros estimados en un monto de S/. 105 668 889. Asimismo, también existen aproximadamente 20 proyectos de este mismo tipo en formulación. El desarrollo de estos proyectos implicaría un desarrollo en el sector de pesca [3].

2.1.2.2 Resultados de la consultoría

De acuerdo con el desarrollo de los objetivos y al alcance de la consultoría se han obtenido cuatro productos que representan los resultados del estudio: Mapeo de clústeres Como arte de los resultados de la consultoría, se han identificado y mapeado 41 clústeres en el Perú [3]. Para ello, se ha realizado un extenso proceso de entrevistas con especialistas de amplio conocimiento sobre la realidad económica de los negocios del Perú, así como con los directores y gerentes de empresas representativas de diferentes sectores. Esto ha permitido contrastar de la mejor manera el análisis previo de fuentes secundarias. Elaboración de bases de datos Se han desarrollado 41 bases de datos para cada uno de los clústeres identificados a partir del análisis del Directorio Nacional de Empresas Manufactureras 2012 de PRODUCE, la base de datos de "Peru The Top 10,000" y de estadísticas de producción y exportaciones del Perú. Las bases de datos contienen información de las empresas identificadas en temas relacionados con la ubicación, número de trabajadores, especialización de las empresas, posición en la cadena de valor del clúster y facturación y exportaciones estimadas. La priorización de clústeres se ha desarrollado y aplicado una metodología para la ponderación y priorización de los clústeres que consta de cinco criterios de priorización: 1) Masa crítica empresarial, 2) potencial de crecimiento del negocio (masa crítica de mercado), 3) ventaja

competitiva del clúster, 4) efecto de arrastre de la cadena en términos de empresas, 5) ocupación y tecnología, que responden a los objetivos y requerimientos del estudio. Como producto de su aplicación, se obtuvo un ranking y priorización referencial de los 41 clústeres identificados. Durante el proceso de priorización, se desarrolló un taller participativo en el cual se enfatizó la diversificación de negocios como parte de la priorización. En cuanto a diagnósticos de clústeres, se han elaborado 17 diagnósticos de los clústeres identificados con la finalidad de obtener una primera caracterización de negocios diversificados. En los diagnósticos se profundizan temas relacionados con la ubicación y delimitación geográfica, dimensión (facturación, exportaciones, número de empresas, etc.), mapeo de la cadena de valor, análisis de las exportaciones e identificación de barreras y retos estratégicos de los clústeres [4]. Cabe mencionar que los diagnósticos no son exhaustivos, porque han sido elaborados con la información obtenida durante el proceso de mapeo de clústeres. Sin embargo, representan un importante insumo para etapas posteriores en la implementación de iniciativas clústeres.

2.2 Redes neuronales artificiales.

Según los autores [5][6], una red neuronal es un modelo computacional con un conjunto de propiedades específicas, consisten en una simulación de las propiedades observadas en los sistemas neuronales biológicos a través de modelos matemáticos recreados mediante mecanismos artificiales (como un circuito integrado, un ordenador o un conjunto de válvulas). Tiene el objetivo: conseguir que las máquinas den respuestas similares a las que es capaz de dar el cerebro, que se caracterizan por su generalización y su robustez. La habilidad de adaptarse o aprender, generalizar u organizar la información, todo ello basado en un procesamiento eminentemente paralelo. Las redes neuronales son modelos que intentan reproducir el comportamiento del cerebro. Los mismos constan de dispositivos elementales de proceso: las neuronas.

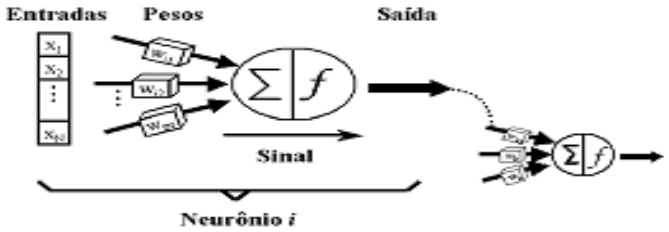


Figura Nro. 3. Entradas, pesos y salida.

2.2.1 Funcionamiento

Una red neuronal se compone de unidades llamadas neuronas. Cada neurona recibe una serie de entradas a través de interconexiones y emite una salida. Esta salida viene dada por tres funciones:

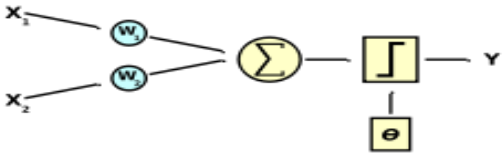


Figura Nro. 4. Salidas de neuronas recibidas por tres funciones.

a) **Una función de propagación (también conocida como función de excitación)**, por lo general consiste en el sumatorio de cada entrada multiplicada por el peso de su interconexión (valor neto). Si el peso es positivo, la conexión se denomina ex citatoria; si es negativo, se denomina inhibitoria.
b) **Una función de activación**, modifica la anterior. Puede no existir, siendo en este caso la salida la misma función de propagación.
c) **Una función de transferencia**, se aplica al valor devuelto por la función de activación. Se utiliza para acotar la salida de la neurona y generalmente viene dada por la interpretación que queramos darle a dichas salidas. Algunas de las más utilizadas son la función sigmoidea (para obtener valores en el intervalo [0,1]) y la tangente hiperbólica (para obtener valores en el intervalo [-1,1]) [7].

2.2.2 Niveles o capas de neuronas

La distribución de neuronas dentro de la red se realiza formando niveles o capas de un número determinado de neuronas cada una. Como se mencionó en el apartado anterior, se pueden distinguir tres tipos de capas [8]:

a) **De entrada:** es la capa que recibe directamente la información proveniente de las fuentes externas a la red
b) **Ocultas:** son internas a la red y no tienen contacto directo con el entorno exterior. Pueden estar interconectadas de distintas maneras, lo que determina junto con su número las distintas tipologías de redes.
c) **De salida:** transfieren información de la red hacia el exterior.

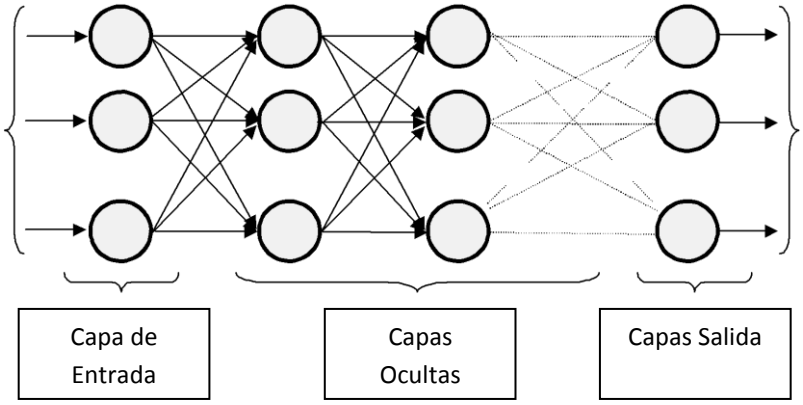


Figura Nro. 5, niveles de capas de neuronas.

La misma está constituida por neuronas interconectadas y arregladas en tres capas (esto último puede variar). Los datos ingresan por medio de la "capa de entrada", pasan a través de la "capa oculta" y salen por la "capa de salida". Cabe mencionar que la capa oculta puede estar constituida por varias capas [6].

Antes de comenzar el estudio sobre las redes neuronales, se debe aprender algo sobre las neuronas y de cómo ellas son utilizadas por una red neuronal. En la Figura N°5 se compara una neurona biológica con una neurona artificial. En la misma se pueden observar las similitudes entre ambas (tienen entradas, utilizan pesos y generan salidas)

Mientras una neurona es muy pequeña en sí misma, cuando se combinan cientos, miles o millones de ellas pueden resolver problemas muy complejos. Por ejemplo, el cerebro humano se compone de billones de tales neuronas [9].

2.2.3 Función de entrada (input function).

La neurona trata a muchos valores de entrada como si fueran uno solo; esto recibe el nombre de entrada global. Por lo tanto, ahora nos enfrentamos al problema de cómo se pueden combinar estas simples entradas (ini1, ini2,...) dentro de la entrada global, gin¡. Esto se logra a través de la función de entrada, la cual se calcula a partir del vector entrada [8].

Los valores de entrada se multiplican por los pesos anteriormente ingresados a la neurona. Por consiguiente, los pesos que generalmente no están restringidos cambian la medida de influencia que tienen los valores de entrada. Es decir, permiten que un gran valor de entrada tenga solamente una pequeña influencia, si estos son lo suficientemente pequeños [9].

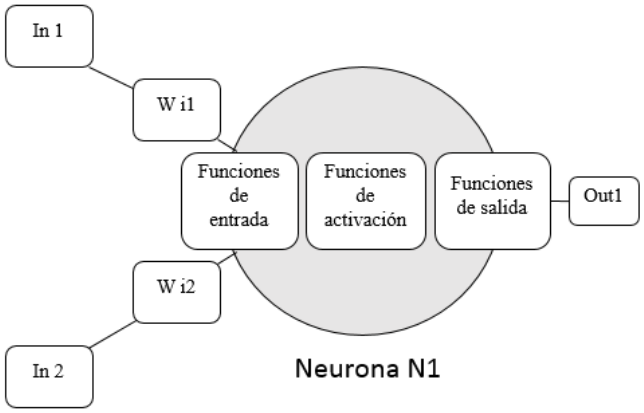


Figura Nro. 6. Función de neurona con 2 entradas y una salida.

2.2.4 Función de salida o transferencia

Existen cuatro funciones de transferencia típicas que determinan distintos tipos de neuronas:

a) Función escalón.
b) Función lineal y mixta.
c) Sigmoidal.
d) Función gaussiana.

La función escalón únicamente se utiliza cuando las salidas de la red son binarias. La salida de una neurona se activa sólo cuando el estado de activación es mayor o igual a cierto valor umbral. La función lineal o identidad equivale a no aplicar función de salida. Las funciones mixtas y sigmoidal son las más apropiadas cuando queremos como salida información analógica [10].

2.2.5 Arquitectura de red

La definición de arquitectura es un punto importante en el modelaje de una red neuronal, porque ella restringe un tipo de problema que puede ser tratado. Por ejemplo, las redes de una capa. Una red también puede estar formada por múltiples capas, las que pueden ser clasificadas en tres grupos: capa de entrada, capas intermediarias u ocultas y capas de salida [11].

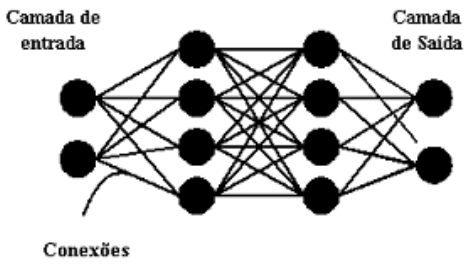


Figura Nro. 7, arquitectura de red.

2.2.6 Redes recurrentes

Redes recurrentes son aquellas que poseen conexiones de realimentación, como son vistas en la figura 8, las cuales proporcionan un comportamiento dinámico. El modelo de Hopfield es un ejemplo de red neuronal recurrente [8].

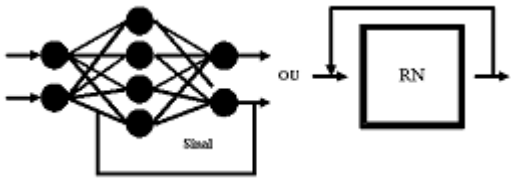


Figura Nro. 8, redes recurrentes.

En general, los siguientes parámetros son importantes para definir la arquitectura de una red neural: **número de capas**, **número de neuronas en cada capa** y **tipo de conexión entre dos neuronas**, que definen la red de feedforward o recurrentes.

2.2.7 Aprendizaje de un RNA

[5] Una propiedad importante de las redes neuronales es la habilidad de aprender a partir de su ambiente. Eso es realizado a través de un proceso interactivo de ajustes aplicado a sus pesos de conexión entre dos neuronas, denominados entrenamiento. Existen muchos algoritmos de aprendizaje. Cada uno sirve para determinar redes neuronales. Entre los principales se tienen:

Aprendizaje por corrección de error. Algoritmo muy conocido basado en la regla Delta, que busca minimizar la función de error usando un gradiente descendente.

Este es el principio usado no algoritmo BackPropagation, muy utilizado, para el entrenamiento de redes de múltiples capas como la Multilayer-Perceptron [12]

Aprendizaje competitivo. La cual dos neuronas de una capa compiten entre sí por el privilegio de permanecer activos, tal que una neurona con mayor actividad será el único que participará del proceso de aprendizaje [13].

Aprendizaje de boltzmann. Es una regla de aprendizaje estocástico obtenido a partir de principios de teórico de información y de termodinámica. El objetivo de aprendizaje de Boltzmann es ajustar los pasos de conexión de tal forma que el estado de las unidades visibles satisfaga una distribución de probabilidades deseada en particular [11].

Aprendizaje supervisado. Se utiliza un agente externo que indica a la red la respuesta deseada para el patrón de entrada.

Refuerzo. Es una variante de aprendizaje supervisado a la cual se informa a la red solamente una crítica de corrección de salida de red y no la respuesta correcta en sí.

Aprendizaje No supervisado (auto-organización). No existe un agente externo indicando la respuesta deseada para los patrones de entrada. Este tipo de aprendizaje es utilizado en los modelos de Mapas de Kohonen.

2.3 Clústering

El objetivo del clústering es reducir la cantidad de datos mediante la caracterización o agrupamiento de datos con características similares. Esta agrupación es acorde con los procesos humanos de información y una de las motivaciones para usar algoritmos clústering es proveer herramientas automáticas que ayuden a la construcción de taxonomías. Los métodos pueden también ser usados para minimizar los efectos de los factores humanos que afectan el proceso de clasificación [14].

Como se puede observar, el problema de clústering consiste en contar una partición óptima en el sentido que los elementos de la partición se entiendan como clases o tipos de datos. De acuerdo con los algoritmos de clústering son métodos para dividir un conjunto de observaciones en grupos de tal manera que miembros del

mismo grupo son más parecidos que miembros de distintos grupos. Para dejar claro el concepto de clústering se dan las siguientes definiciones [15].

2.3.1 Algoritmo clústeres de microsoft (Analysis Services - Minería de datos).

El algoritmo de clústeres de microsoft es un algoritmo de segmentación suministrado por SQL Server 2008 Analysis Services (SSAS). El algoritmo utiliza técnicas iterativas, para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares [16]. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones [17].

Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual.

Por ejemplo, puede discernir lógicamente que las personas que se desplazan a sus trabajos en bicicleta no viven, por lo general, a gran distancia de sus centros de trabajo. Sin embargo, el algoritmo puede encontrar otras características que no son evidentes acerca de los trabajadores que se desplazan en bicicleta. En el siguiente diagrama, el clúster A representa los datos sobre las personas que suelen conducir hasta el trabajo, en tanto que el clúster B representa los datos sobre las personas que van hasta allí en bicicleta.

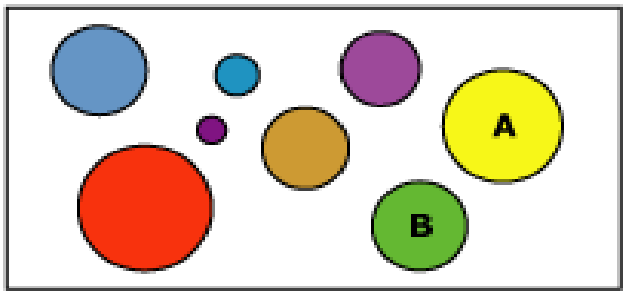


Figura Nro. 9, Algoritmo clústeres de microsoft (Analysis Services - Minería de datos).

El algoritmo de agrupación en clústeres se diferencia de otros algoritmos de minería de datos, en que no se tiene que designar una columna de predicción para

generar un modelo de agrupación en clústeres. El algoritmo de agrupación en clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo [6].

2.3.2 Cómo funciona el algoritmo

El algoritmo de agrupación en clústeres de microsoft identifica primero las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas. Un gráfico de dispersión es una forma útil de representar visualmente el modo como el algoritmo agrupa los datos, tal como se muestra en el siguiente diagrama. El gráfico de dispersión representa todos los casos del conjunto de datos; cada caso es un punto del gráfico. Los clústeres agrupan los puntos del gráfico e ilustran las relaciones que identifica el algoritmo [5].

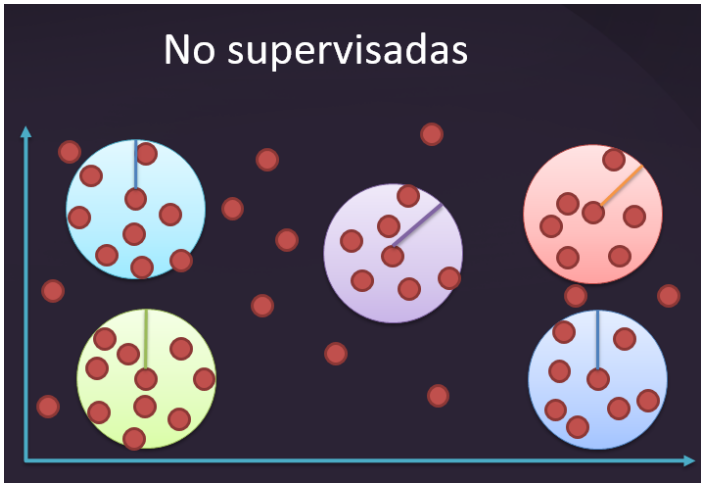


Figura Nro. 10, funcionamiento del algoritmo.

Después de definir los clústeres, el algoritmo calcula el grado de perfección con que los clústeres representan las agrupaciones de puntos y, a continuación, intenta volver a definir las agrupaciones para crear clústeres que representen mejor los datos. El algoritmo establece una iteración en este proceso hasta que ya no es posible mejorar los resultados mediante la redefinición de los clústeres [18].

Puede personalizar el funcionamiento del algoritmo seleccionando una técnica de agrupación en clústeres, limitando el número máximo de clústeres o cambiando la cantidad de soporte que se requiere para crear un clúster.

2.3.3 Datos requeridos para los modelos de agrupación en clústeres

Al preparar los datos para su uso en el entrenamiento de un modelo de agrupación en clústeres, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma como estos datos se utilizan. Según microsoft, los requisitos para un modelo de agrupación en clústeres son los siguientes:

a) **Una única columna key.** Cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
b) **Columnas de entrada.** Cada modelo debe tener al menos una columna de entrada que contenga los valores que se utilizan para generar los clústeres. Puede tener tantas columnas de entrada como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el modelo.
c) **Una columna de predicción opcional.** El algoritmo no necesita una columna de predicción para generar el modelo, pero puede agregar una columna de predicción de casi cualquier tipo de datos. Los valores de la columna de predicción se pueden tratar como entradas del modelo de agrupación en clústeres, o se puede especificar que sólo se utilicen para las predicciones. Por ejemplo, si desea predecir los ingresos del cliente agrupando en clústeres de acuerdo con datos demográficos como la región o la edad, se deben especificar los ingresos como **PredictOnly** y agregar todas las demás columnas, como la región o la edad, como entradas [2].

2.3.4 Modelo de agrupación en clústeres.

Para explorar el modelo, puede utilizar el **Visor de clústeres de Microsoft**. Cuando se observa un modelo de agrupación en clústeres, el Analysis Services presenta los clústeres en un diagrama que muestra las relaciones existentes entre ellos, además de un perfil detallado de cada clúster, una lista de los atributos que

diferencian cada clúster de los demás, y las características de todo el conjunto de datos de entrenamiento.

2.3.5 **Métodos para crear clústeres de microsoft**

El algoritmo de clústeres de Microsoft proporciona dos métodos para crear clústeres y asignar puntos de datos a dichos clústeres. El primero, el algoritmo mediana-K es un método de agrupación en clústeres duro. Esto significa que un punto de datos puede pertenecer a un solo clúster, únicamente se calcula una probabilidad de pertenencia de cada punto de datos de ese clúster. El segundo, el método Expectation Maximization (EM), es un método de agrupación en clústeres blando. Esto significa que un punto de datos siempre pertenece a varios clústeres, y que se calcula una probabilidad para cada combinación de punto de datos y clúster [17].

Puede elegir el algoritmo que desee utilizar estableciendo el parámetro CLUSTERING_METHOD. El método predeterminado para la agrupación en clústeres es el método EM escalable.

2.3.6 Agrupación en clústeres EM.

En el método de agrupación en clústeres EM, el algoritmo refina de forma iterativa un modelo de clústeres inicial para ajustar los datos y determina la probabilidad de que un punto de datos exista en un clúster. El algoritmo finaliza el proceso cuando el modelo probabilístico ajusta los datos. La función utilizada para determinar el ajuste es el logaritmo de la probabilidad de los datos dado el modelo [19].

Los resultados del método de agrupación en clústeres EM son probabilísticos. Esto significa que cada punto de datos pertenece a todos los clústeres, pero cada asignación de un punto de datos a un clúster tiene una probabilidad diferente. Dado que el método permite que los clústeres se superpongan, la suma de los elementos de todos los clústeres puede superar la totalidad de los elementos existentes en el

conjunto de entrenamiento. En los resultados del modelo de minería de datos, las puntuaciones que indican soporte se ajustan para tener en cuenta este hecho [10]

El algoritmo EM es el algoritmo predeterminado utilizado en los modelos de agrupación en clústeres de Microsoft. Este algoritmo se utiliza como algoritmo predeterminado porque proporciona numerosas ventajas comparado con la agrupación en clústeres mediana-K:

a) Requiere examinar la base de datos, máximo una vez.
b) Funciona incluso si la cantidad de memoria (RAM) es limitada.
c) Tiene la capacidad de utilizar un cursor de sólo avance.
d) Sus resultados superan los obtenidos por los métodos de muestreo

a) Agrupación en clústeres mediana-K

La agrupación en clústeres mediana-K es un método muy conocido para asignar la pertenencia al clúster que consiste en minimizar las diferencias entre los elementos de un clúster al tiempo que se maximiza la distancia entre los clústeres [17]. El término "mediana" hace referencia al centroide del clúster, que es un punto de datos que se elige arbitrariamente y que se refina de forma iterativa hasta que representa la verdadera media de todos los puntos de datos del clúster. La "K" hace referencia a un número arbitrario de puntos que se utilizan para inicializar el proceso de agrupación en clústeres [18]. El algoritmo mediana-K calcula las distancias euclidianas cuadradas entre los registros de datos de un clúster y el vector que representa la media de clústeres, y converge en un conjunto final de K clústeres cuando la suma alcanza su valor mínimo.

El algoritmo mediana-K asigna cada punto de datos a un solo clúster y no permite la incertidumbre en la pertenencia. En un clúster, la pertenencia se expresa como una distancia desde el centroide [20].

Normalmente, el algoritmo mediana-K se utiliza para crear clústeres de atributos continuos, donde el cálculo de la distancia a una media se realiza de manera sencilla. Sin embargo, la implementación de Microsoft adapta el método mediana-K a atributos discretos de clúster mediante el uso de probabilidades.

El algoritmo mediana-K proporciona dos métodos para realizar un muestreo en el conjunto de datos: mediana-K no escalable, que carga el conjunto de datos completo y realiza una pasada de agrupación en clústeres, y mediana-K escalable, donde el algoritmo utiliza los primeros 50.000 casos y lee más casos únicamente si necesita más datos para lograr un buen ajuste del modelo a los datos.

2.3.7 El algoritmo de distancia euclidiana

2.3.7.1 Distancia euclidiana

La distancia euclídea es la disimilaridad más conocida y más sencilla de comprender, pues su definición coincide con el concepto más común de distancia.

Su expresión es la siguiente: $d(i,j) = (W_i - W_j)' (W_i - W_j)$

La distancia euclídea, a pesar de su sencillez de cálculo y de que verifica algunas propiedades interesantes, tiene dos graves inconvenientes:

a) El primero de ellos es que la euclídea es una distancia sensible a las unidades de medida de las variables: las diferencias entre los valores de variables medidas con valores altos contribuirán en mucha mayor medida de que las diferencias entre los valores de las variables medidas con valores bajos. Como consecuencia de ello, los cambios de escala determinarán, también, cambios en la distancia entre los individuos. Una posible vía de solución de este problema es la tipificación previa de las variables, o la utilización de la distancia euclídea normalizada.
b) El segundo inconveniente no se deriva directamente de la utilización de este tipo de distancia, sino de la naturaleza de las variables. Si las variables utilizadas están correlacionadas, estas variables nos darán una información, en gran medida redundante. Parte de las diferencias entre los valores individuales de algunas variables podrían explicarse por las diferencias en otras variables. Como consecuencia de ello la distancia euclídea inflará la disimilaridad o divergencia entre los individuos.

La solución a este problema pasa por analizar las componentes principales (que están incorrelacionadas) en vez de las variables originales. Otra posible solución es ponderar la contribución de cada par de variables con pesos inversamente

proporcionales a las correlaciones, lo que nos lleva, como veremos, a la utilización de la distancia de Mahalanobis.

La distancia euclídea será, en consecuencia, recomendable cuando las variables sean homogéneas y estén medidas en unidades similares y/o cuando se desconozca la matriz de varianzas.

2.4 Elección de la metodología

La metodología SEMMA y la CRISP-DM desarrollan el proyecto Data Mining en distintas fases que se encuentran interrelacionadas llegando a producir un desarrollo iterativo e interactivo.

Una gran diferencia se encuentra en el comienzo del proyecto que será analizado, porque para la metodología SEMMA se centra más en las características técnicas en cada proceso, donde empieza con el muestreo de los datos; en cambio, la metodología CRISP -DM realiza la minería de datos, de manera más amplia de acuerdo con los objetivos empresariales y el análisis del problema empresarial para su transformación en un problema técnico. De manera global se puede considerar que la metodología CRISP-DM es más cercana al concepto real de cada proyecto, porque se integrara con una metodología de gestión de proyectos específica que completaría las tareas administrativas y técnicas. También se encuentra una gran diferencia en el tema comercial. Donde SEMMA es ligada a los productos SAS, permitiéndole ser abierta a sus aspectos generales, donde se encuentra implementada. Por otro lado, la metodología CRISP-DM ha sido diseñada como metodología neutra, con distribución libre y gratuita.

Para poder elegir la metodología a seguir durante todo el proyecto es necesario revisar las diferencias y semejanzas que existen.

A continuación en la tabla 2 se presenta un cuadro comparativo entre la metodología de CRISP-DM y SEMMA.

Tabla 2 - Cuadro comparativo Crisp-DM vs SEMMA.

Cuadro comparativo de Crisp-DM Vs SEMMA		
Objetivo	**Crisp-DM**	**SEMMA**
Permite elección libre de las herramientas	SI	NO

Cantidad de fases	6	5
Todas las fases pueden relacionarse	SI	NO
Considera los motivos del proyecto	NO	NO
Considera la naturaleza del interés de las partes	NO	NO
Considera otras partes no técnicos	SI	NO
Identifica claramente las variables sobre las cuales el proyecto tiene impacto	NO	NO
Esta detallada paso a paso cada etapa del método	NO	NO
Identifica problemas de inteligencia de negocio (PIN)	SI	NO
Identifica una caracterización abstracta de PIN	NO	Parcialmente.
Identifica técnicas de exploración de información (TEI) utilizables	SI	SI
Identifica relaciones entre las TEI y los PIN	NO	Parcialmente.
Identifica procesos de explotación de información (procesos PINxTEI)	NO	Parcialmente

La metodología SEMMA se centra más en las características técnicas en cada proceso, donde empieza con el muestreo de los datos; en cambio, la metodología CRISP -DM realiza la minería de datos, de manera más amplia de acuerdo con los objetivos empresariales y el análisis del problema empresarial para su transformación en un problema técnico. De manera global se puede considerar que la metodología CRISP-DM es más cercana al concepto real de cada proyecto, por tal razón se elige la metodología CRISP -DM.

2.5 La metodología de Crisp-DM

La elección de este estándar incluye un modelo y una guía, estructurados en **seis fases**, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores [21].

Basado en este análisis, la mejor opción para este proyecto sería CRISP-DM ubicando las características del cliente, finanzas alumnos de la Universidad Peruana Unión.

Para implementar una tecnología en un negocio, se requiere una metodología. La mayoría de las consultoras especializadas en alguna tecnología tienen, por lo menos, una metodología, según los tipos de proyectos que aborden. Estos métodos son definidos a partir de sus experiencias y tomando lo mejor de los procedimientos más exitosos o populares. Contar con una metodología,

se ha convertido tan importante y necesario como la carta de presentación de las empresas.

La metodología de trabajo utilizada, para lograr el objetivo planteado, fue La CRISP-DM, la cual consiste en un conjunto de tareas descritas en cuatro niveles de abstracción: fase, tarea genérica, tarea especializada, e instancia de proceso, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos [22].

El estándar incluye un modelo y una guía, estructurados en seis fases, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores.

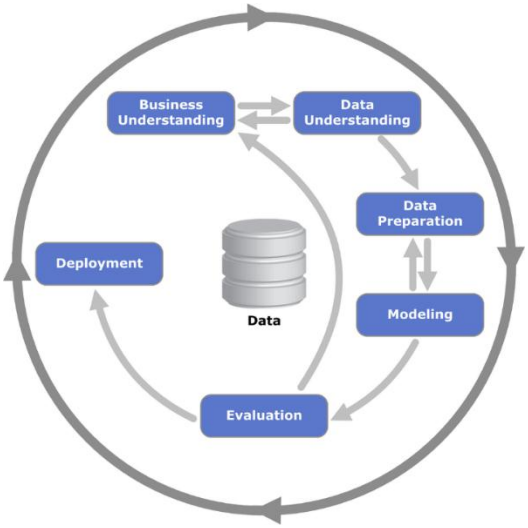


Figura Nro. 11, modelo de la metodologia de CRISP-DM.

El estándar incluye un modelo y una guía, estructurados en seis fases, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores.

Comprensión del negocio (Objetivos y requerimientos desde una perspectiva no técnica)

a) Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)
b) Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio)

c) Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
d) Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

Comprensión de los datos (Familiarizarse con los datos teniendo presente los objetivos del negocio)

a) Recopilación inicial de datos
b) Descripción de los datos
c) Exploración de los datos
d) Verificación de calidad de datos

Preparación de los datos (Obtener la vista minable o dataset)

a) Selección de los datos
b) Limpieza de datos
c) Construcción de datos
d) Integración de datos
e) Formateo de datos

Modelado (Aplicar las técnicas de minería de datos a los dataset)

a) Selección de la técnica de modelado
b) Diseño de la evaluación
c) Construcción del modelo
d) Evaluación del modelo

Evaluación (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio)

a) Evaluación de resultados
b) Revisar el proceso
c) Establecimiento de los siguientes pasos o acciones

Despliegue (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización)

a) Planificación de despliegue

b) Planificación de la monitorización y del mantenimiento
c) Generación de informe final
d) Revisión del proyecto

2.5.1 Procesos de desarrollo con CRISP-DM

El modelo de proceso corriente para la minería de datos proporciona una descripción del ciclo de vida del proyecto de minería de datos. Este contiene las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción, no es posible identificar todas las relaciones. Las relaciones podrían existir entre cualquier tarea de minería de datos según los objetivos, el contexto, y lo más importante el interés del usuario sobre los datos [23]. El movimiento hacia adelante y hacia atrás entre fases diferentes es siempre requerido.

El círculo externo en la Figura 17 simboliza la naturaleza cíclica de la minería de datos. La minería de datos no se termina una vez que la solución es desplegada. Las informaciones ocultas durante el proceso y la solución desplegada pueden provocar nuevas, a menudo más preguntas enfocadas en el negocio. En el siguiente, brevemente perfilamos cada fase:

a) Comprensión del negocio. Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos [24].

b) Comprensión de los datos. La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta [21].

c) Preparación de datos. La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [los datos que serán provistos en las herramientas de modelado] de los datos en brutos iniciales. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan [21].

d) Modelado. En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario [25].

e) Evaluación. En esta etapa del proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos. Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida [26].

f) Desarrollo. La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo como el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización; por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos de mercadeo [22]. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. [10].

La figura 12 presenta un contexto de fases acompañadas por tareas genéricas y las salidas. En las secciones siguientes, describimos cada tarea genérica y sus salidas más detalladamente. Enfocamos nuestra atención en descripciones de tarea y los resúmenes de salidas.

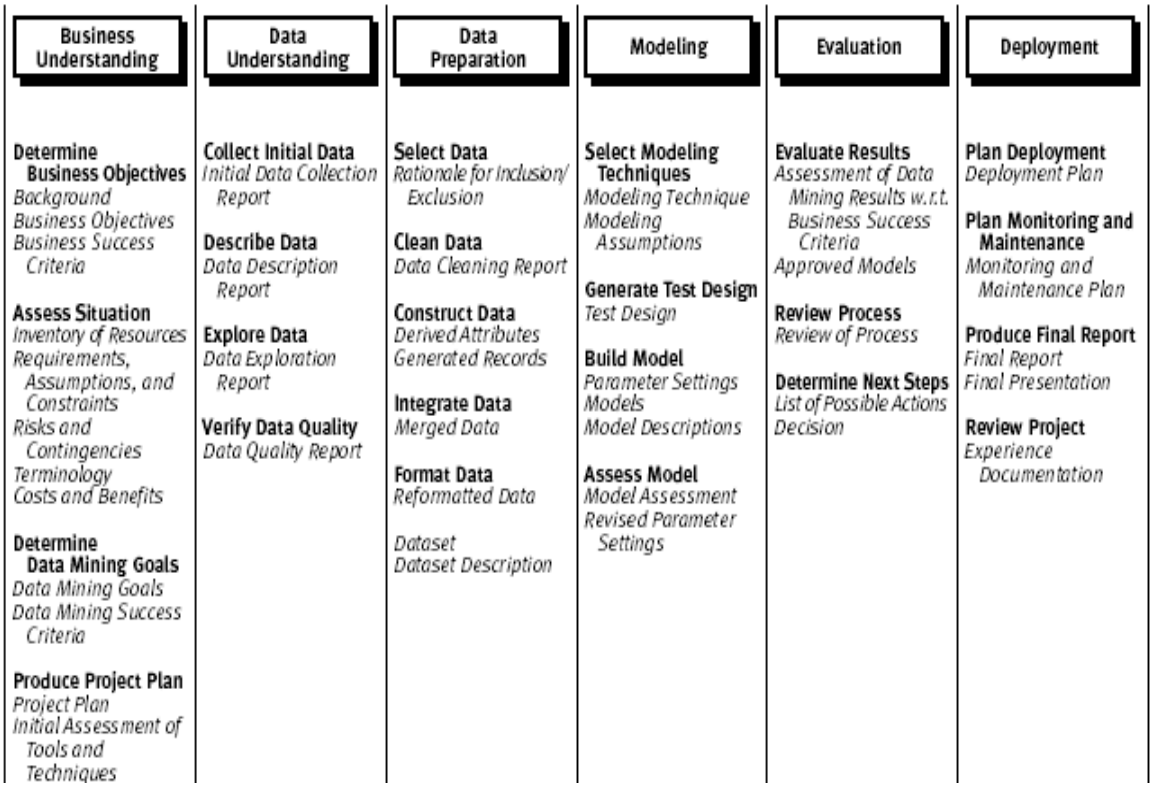


Figura Nro. 12, procesos de desarrollo con CRISP-DM.

2.5.2 Aplicación de CRISP-DM en el caso de estudio

2.5.2.1 Comprensión del negocio.

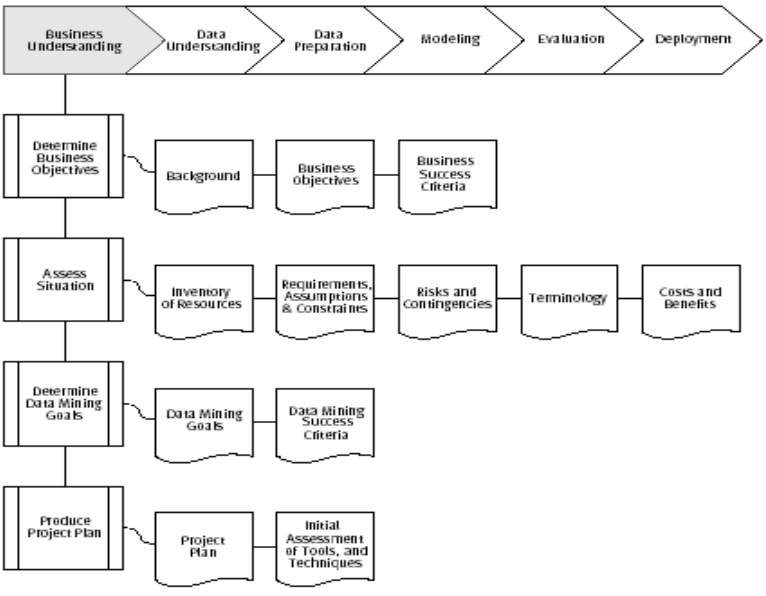


Figura Nro. 13, comprension del negocio.

2.5.2.1.1 Determinación de objetivos de negocio

a) **Tarea: Determinar los objetivos de negocio.** El primer objetivo del analista de datos para un contexto, es entender, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. A menudo, el cliente tiene muchos objetivos que compiten y restricciones que deben ser correctamente equilibrados. El objetivo del analista es mostrar (destapar) factores importantes, en el principio, esto puede influir en el resultado del proyecto. Una consecuencia posible de la negligencia de este paso es gastar un gran esfuerzo produciendo respuestas correctas a preguntas incorrectas o erradas [27].

b) **Salida: Contexto.** Registre la información que conoce sobre la situación de negocio de la organización en el principio del proyecto.

c) **Objetivos de negocio**. Es el objetivo primario del cliente, desde una perspectiva de negocio. Además de los objetivos del negocio primario, allí hay típicamente otras preguntas de negocio relacionadas con lo que al cliente le gustaría administrar. Por ejemplo, el objetivo primario de negocio podría ser mantener a clientes corrientes por predicción cuando ellos son propensos a moverse a un competidor.

d) **Criterios de éxito de negocio**. Describa los criterios para un resultado acertado o útil al proyecto desde el punto de vista del negocio. Esto podría ser bastante específico y capaz de ser medido objetivamente; por ejemplo, la reducción de clientes se revuelve a un cierto nivel o valor, o esto podría ser general y subjetivo, como "dar ideas útiles en las relaciones". En este último caso, esto debería indicarse quien hace el juicio subjetivo [23].

2.5.2.1.2 Evaluación de la situación

a) **Tarea: Evaluar la situación.** Esta tarea implica la investigación más detallada sobre todos los recursos, restricciones, presunciones, y otros factores que deberían ser considerados en la determinación del objetivo de análisis de datos y el plan de proyecto. En la tarea anterior, su objetivo es para ponerse rápidamente en el quid de la situación. Aquí usted quiere ampliarse sobre los detalles.

b) **Salida: Inventario de recursos.** Listar los recursos disponibles para el proyecto, incluyendo el personal (expertos de negocio, expertos de datos, soportes técnicos, expertos en minería de datos), datos (extractos fijos, aproximaciones a la vida, almacenes de datos, u datos operacionales), recursos computacionales

(plataformas de hardware) y software (herramientas de minería de datos, otros software relevantes).

c) **Requerimientos, presunciones, y restricciones**. Listar todos los requerimientos del proyecto, incluyendo el programa de terminación, la comprensibilidad y calidad de los resultados, y la seguridad, así como las cuestiones legales. Como parte de esta salida, asegúrese que le permitan usar los datos [24]. Listar las presunciones hechas por el proyecto. Estas pueden ser presunciones sobre los datos que pueden ser verificados durante la minería de datos, también puede incluir presunciones no-comprobables sobre el negocio relacionado con el proyecto. Es en particular importante listar si esto afectará la validez de los resultados.

Listar las restricciones sobre el proyecto. Estas pueden ser restricciones sobre la disponibilidad de recursos, puede también incluir coacciones tecnológicas como el tamaño de conjunto de datos lo que es práctico para usar el modelado.

d) Riesgos y contingencias. Listar los riesgos o los acontecimientos que podrían retrasar el proyecto o hacer que ello falle. Listar los planes de contingencia correspondientes, que acción será tomada si estos riesgos o acontecimientos ocurren.

e) **Terminología.** Compile un glosario de terminología relevante al proyecto. Esto puede incluir dos componentes. Un glosario de terminología relevante del negocio, forma la parte de la comprensión del negocio disponible al proyecto. La construcción de este glosario es una útil "evocación al conocimiento" y un ejercicio de educación.

f) **Costos y beneficios**. Construya un análisis de costo-beneficio para el proyecto, que compare los gastos del proyecto con los beneficios potenciales al negocio si esto es exitoso. La comparación debería ser tan específica como posible. Por ejemplo, use medidas monetarias en una situación comercial [25].

2.5.2.1.3 Determinación de los objetivos de la minería de datos

a) Tarea: Determinar los objetivos de la minería de datos. Un objetivo de negocio declara objetivos en la terminología de negocio. Un objetivo de minería de datos declara objetivos de proyecto en términos técnicos [27]. Por ejemplo, el objetivo de negocio podría ser "Aumentar catálogos de ventas a clientes existentes." Un objetivo de minería de datos podrían ser "Predecir cuantas

baratijas un cliente comprará, obteniendo datos de sus compras de tres años pasados, información demográfica (edad, sueldo, ciudad, etc.), y el precio del artículo."

b) **Salida: Objetivos de la minería de datos**. Describir las salidas intencionadas del proyecto que permiten el logro de los objetivos de negocio.

c) **Criterios de éxito de la minería de datos**. Definir los criterios de un resultado exitoso para el proyecto en términos técnicos -por ejemplo, un cierto nivel de predicción precisa o un perfil de inclinación-a-comprar con un determinado grado de "elevación". Como con un criterio de éxito de negocio, puede ser necesario describir estos en términos subjetivos, en este caso la persona o las personas que hacen el juicio subjetivo deberían ser identificadas.

2.5.2.1.4 Producir el plan del proyecto

a) Tarea: Producir el plan del proyecto. Describir el plan intencionado para alcanzar los objetivos de minería de datos y así alcanzar los objetivos de negocio. El plan debería especificar los pasos para ser realizados durante el resto del proyecto, incluyendo la selección inicial de herramientas y técnicas.

b) Salida: Plan del Proyecto. Listar las etapas a ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas y dependencias. Donde sea posible, haga explícito las iteraciones en gran escala en el proceso de minería de datos; por ejemplo, las repeticiones del modelado y las fases de evaluación. Como parte del plan de proyecto, es también importante analizar dependencias entre la planificación de tiempo y los riesgos. El plan de proyecto es un documento dinámico en el sentido de que en el final de cada fase, son necesarios una revisión del progreso y los logros, asimismo una actualización correspondiente del plan de proyecto es recomendable [22].

c) **Evaluación inicial de herramientas y técnicas**. En la final de la primera fase, una evaluación inicial de herramientas y técnicas debería ser realizada. Aquí, por ejemplo, usted selecciona una herramienta de minería de datos que soporte varios métodos para las distintas etapas del proceso. Es importante evaluar herramientas y técnicas temprano en el proceso desde la selección de herramientas y técnicas y esto puede influir en el proyecto entero.

2.5.2.2 Comprensión de datos

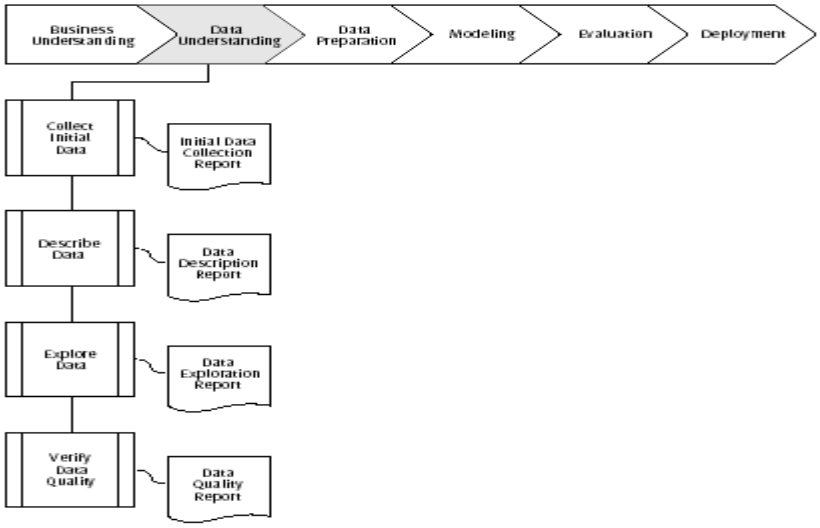


Figura Nro. 14, comprension de datos.

2.5.2.2.1 Recolección de datos iniciales

a) Tarea: Recolectar datos iniciales. Adquiera en el proyecto los datos (o el acceso a los datos) listados en los recursos del proyecto. Esta colección inicial incluye carga de datos, si es necesario para la comprensión de los datos.

b) **Salida: Informe de colección de datos inicial.** Liste el conjunto de dato(s) adquirido(s), juntos con sus posiciones, los métodos usados para adquirirlos, y algunos de los problemas encontrados. Registre los problemas encontrados y algunas de las resoluciones alcanzadas. Esto ayudará con la réplica (observación) futura de este proyecto o con la ejecución de proyectos similares futuros [25].

2.5.2.2.2 Describir los datos

a) Tarea: Describir los datos. Examine las propiedades "gruesas" o "superficiales" de los datos e informe adquiridos en los resultados.

b) **Salida: Informe de descripción de datos**. Describa los datos que han sido adquiridos, incluyendo el formato de los datos, la cantidad de datos (por ejemplo, el número de registros y campos en cada tabla), los identificadores de los campos, y cualquier otro rasgo superficial que ha sido descubierto. Evalúe si los datos adquiridos satisfacen las exigencias relevantes.

2.5.2.2.3 Explorar los datos

a) Tarea: Explorar los datos. Esta tarea dirige interrogantes de minería de datos usando preguntas, visualización, y técnicas de reporte. Estos incluyen la

distribución de atributos claves (por ejemplo, el atributo objetivo de una tarea de predicción) relacionados entre pares o pequeños números de atributos, los resultados de simples agregaciones, las propiedades de las subpoblaciones significativas, y análisis estadísticos simples.

b) **Salida: Informe de exploración de datos.** Describa los resultados de esta tarea, incluyendo primeras conclusiones o hipótesis iniciales y su impacto sobre el resto del proyecto. Si es apropiado, incluya gráficos y plots para indicar las características de datos que sugieren más examen de subconjuntos de datos interesantes [21].

2.5.2.2.4 Verificar la calidad de los datos

a) Tarea: Verificar la calidad de los datos. Examine la calidad de los datos, dirigiendo preguntas como: ¿Los datos están completos? (¿Esto cubre todo los casos requeridos)? ¿Son correctos, o estos contienen errores y, si hay errores, que tan comunes son estos? ¿Hay valores omitidos en los datos? Si es así, ¿cómo se representan estos, donde ocurre esto, y que tan comunes son estos?

b) **Salida: Informe de calidad de datos.** Liste los resultados de la verificación de calidad de datos; si existen problemas de calidad, liste las posibles soluciones. Las soluciones a los problemas de calidad de datos generalmente dependen tanto del conocimiento de los datos y como del negocio.

2.5.2.3 Preparación de datos

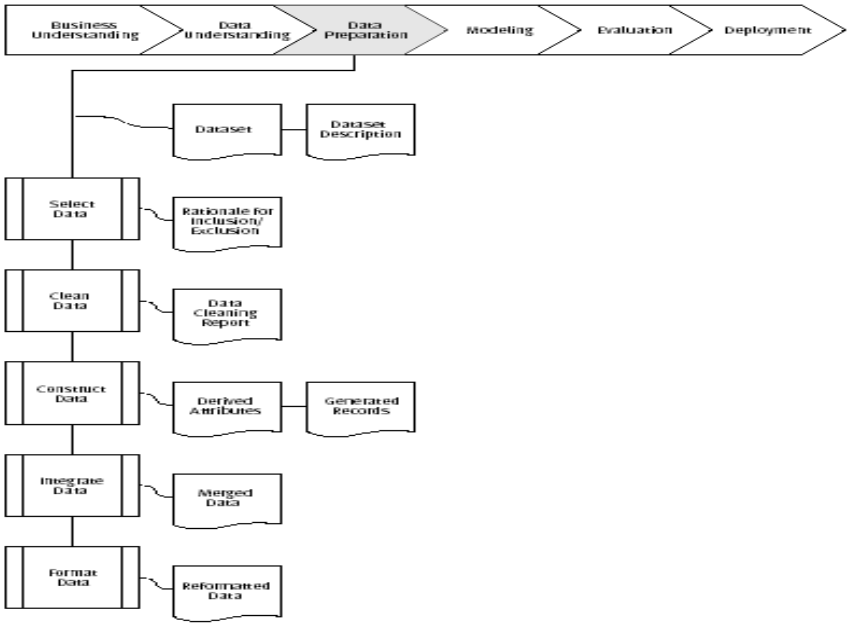


Figura Nro. 15, preparación de datos.

a) **Salida: Conjunto de datos.** Este es el conjunto (o conjuntos) producido por la fase de preparación de datos, que será usada para modelar o para el trabajo principal de análisis del proyecto.

b) **Descripción del conjunto de datos.** Describir el conjunto de dato (o conjuntos) que será usado para el modelado y el trabajo principal de análisis del proyecto.

2.5.2.3.1 Selección de datos

a) **Tarea: Selección de datos.** Decidir qué datos serán usados para el análisis. Los criterios incluyen la importancia a los objetivos de la minería de datos, la calidad, y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos. Note que la selección de datos cubre la selección de atributos (columnas) así como la selección de registros (filas) en una tabla [26].

b) **Salida: Razonamiento para la inclusión/exclusión.** Listar los datos para ser incluidos/excluidos y los motivos para estas decisiones.

2.5.2.3.2 Limpieza de datos

a) **Tarea: Limpiar datos.** Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de los subconjuntos de datos limpios, la inserción de datos por defectos adecuados, o técnicas más ambiciosas tales como la estimación de datos faltantes mediante modelado.

b) **Salida: Informe de la limpieza de los datos.** Describa que decisiones y acciones fueron tomadas para dirigir los problemas de calidad de datos informados durante la tarea de Verificación de Calidad de Datos de los Datos de la fase de Comprensión de Datos. Las transformaciones de los datos para una apropiada limpieza y el posible impacto en el análisis de resultados deberían ser considerados [27].

2.5.2.3.3 Construir datos

a) **Tarea: Construir datos.** Esta tarea incluye la construcción de operaciones de preparación de datos: la producción de atributos derivados, el ingreso de nuevos registros, la transformación de valores para atributos existentes.

b) **Salidas Atributos derivados**. Los atributos derivados son los atributos nuevos que son construidos de uno o más atributos existentes en el mismo registro. Ejemplo: área = longitud * anchura.

c) **Registros generados.** Describa la creación de registros completamente nuevos. Ejemplo, crear registros para los clientes quienes no hicieron compras durante el año pasado. No había ninguna razón de tener tales registros en los datos brutos, pero para el objetivo del modelado esto podría tener sentido para representar explícitamente el hecho que ciertos clientes no hayan hecho compra nada.

2.5.2.3.4 Integrar datos

a) **Tarea Integrar datos.** Estos son los métodos por los cuales la información es combinada de múltiples tablas o registros para crear nuevos registros o valores.

b) **Salida Combinación de datos.** La combinación de tablas se refiere a la unión simultánea de dos o más tablas que tienen información diferente sobre el mismo objeto. Ejemplo, una cadena de venta al público tiene una tabla con la información sobre las características generales de cada tienda (Por ejemplo, el espacio, el tipo de comercio), otra tabla con datos resumidos de las ventas (por ejemplo, el beneficio, el cambio porcentual en ventas desde el año anterior), y el otro con información sobre los datos demográficos del área circundante. Cada una de estas tablas contiene un registro para cada tienda. Estas tablas pueden ser combinadas simultáneamente en una nueva tabla con un registro para cada tienda, combinando campos de las tablas fuentes [25].

2.5.2.3.5 Formatear datos

a) **Tarea Formatear datos.** Formateando transformaciones se refiere a modificaciones principalmente sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.

b) **Salida Datos reformateados.** Algunas herramientas tienen requerimientos sobre el orden de los atributos, tales como el primer campo que es un único identificador para cada registro o el último campo es el campo resultado que el modelo debe predecir. Podría ser importante cambiar el orden de los registros en el conjunto de datos. Quizás la herramienta de modelado requiere que los registros sean clasificados según el valor del atributo de resultado. Comúnmente, los registros del conjunto de datos son ordenados al principio de algún modo, pero el algoritmo que modela necesita que ellos estén en un orden moderadamente arbitrario. [24].

2.5.2.4 Modelado

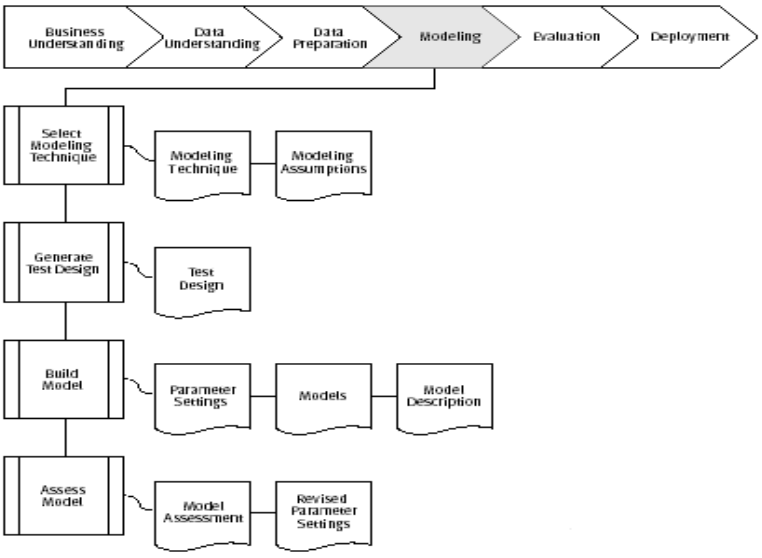


Figura Nro. 16, modelado.

2.5.2.4.1 Selección de la técnica de modelado

a) Tarea: Escoger la técnica de modelado. Como primer paso en modelado, seleccionar la técnica de modelado real que está por ser usado. Aunque se haya podido seleccionar una herramienta durante la fase de comprensión del negocio, esta tarea se refiere a la técnica de modelado específico; por ejemplo, un árbol decisión construido con C4.5, o la generación de red neuronales Back-Propagación. Si múltiples técnicas son aplicadas, se realizan esta tarea separadamente para cada técnica [25].

b) **Salida: Técnicas de modelado**. Se debe documentar la técnica de modelado real que está por ser usado.

c) **Presunciones del modelado.** Muchas técnicas de modelado hacen presunciones específicas sobre los datos; por ejemplo, todos los atributos tengan distribuciones uniformes, no encontrar valores no permitidos, el atributo de clase debe ser simbólico, etc. Registrar cualquiera de tales presunciones hechas [26].

2.5.2.4.2 Generación de la prueba de diseño

a) Tarea: Generar la prueba de diseño. Antes de que nosotros en realidad construyamos un modelo, tenemos que generar un procedimiento o el mecanismo para probar la calidad y validez del modelo. Por ejemplo, en tareas de minería de datos supervisados como la clasificación, esto es común usar tasas de errores como medida de calidad para modelos de minería de datos. Por lo tanto, típicamente se debe separar el conjunto de datos en una serie y en un conjunto de

prueba, construir el modelo sobre el conjunto de series, y estimar su calidad sobre el conjunto de prueba separado [22].

b) **Salida: Prueba de diseño.** Describir el plan intencionado para el entrenamiento, la prueba, y la evaluación de los modelos. Un componente primario del plan determina como dividir un conjunto de datos disponible en datos de entrenamiento, datos de prueba, y conjunto de datos de validación.

2.5.2.4.3 Construcción del modelo

a) Tarea: Construir el modelo. Ejecutar la herramienta de modelado sobre el conjunto de datos preparados para crear uno o más modelos.

b) **Salidas: Parámetro de ajustes**. Con cualquier herramienta de modelado, hay a menudo un gran número de parámetros que pueden ser ajustados. Se debe listar los parámetros y sus valores escogidos, también con el razonamiento para elegir los parámetros de ajustes.

c) Modelos. Estos son los modelos reales producidos por la herramienta de modelado, no un informe.

d) **Descripciones del modelo.** Describir los modelos obtenidos. Informar sobre la interpretación de los modelos y documentar cualquier dificultad encontrada con sus significados.

2.5.2.4.4 Evaluación del modelo

a) Tarea: Evaluar el modelo. El ingeniero de minería de datos interpreta los modelos según su conocimiento de dominio, los criterios de éxitos de minería de datos, y el diseño de prueba deseado. El ingeniero de minería de datos juzga el éxito de la aplicación del modelado y descubre técnicas más técnicamente; él se pone en contacto con analistas de negocio y expertos en el dominio luego para hablar de los resultados de la minería de datos en el contexto de negocio. El ingeniero de minería de datos intenta clasificar los modelos. Él evalúa los modelos según los criterios de evaluación. Tanto como es posible, él también tiene en cuenta objetivos del negocio y criterios de éxito de negocio. [25].

b) **Salida: Evaluación de modelos**. Resumir los resultados de esta tarea, listar las calidades de los modelos generados (por ejemplo, en términos de exactitud), y clasificar su calidad en relación con cada otro.

c) **Parámetros de ajustes revisados.** Según la evaluación del modelo, se debe revisar los parámetros de ajuste y témplelos para la siguiente corrida en la tarea

de construcción del modelo. Repetir la construcción y evaluación del modelo hasta que crea que usted ha encontrado el/los mejor/es modelo/s. Documentar todo como las revisiones y las evaluaciones.

2.5.2.5 Evaluación

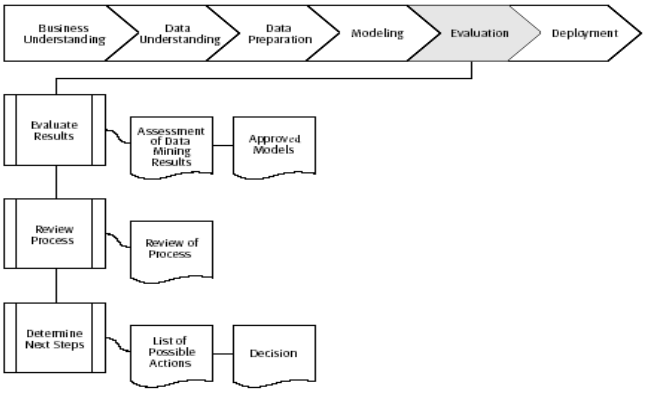


Figura Nro. 17, evaluación.

2.5.2.5.1 Evaluación de los resultados

a) Tarea: Evaluar los resultados. Los pasos de la evaluación anterior trata con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el que este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten.

Además, la evaluación también verifica otros resultados generados por la minería de datos. Los resultados de la minería de datos implican modelos que necesariamente son relacionados con los objetivos originales de negocio y todas los otros descubrimientos que no son relacionados necesariamente con los objetivos originales de negocio, pero también podría revelar desafíos adicionales, información, o insinuaciones para futuras direcciones.

b) **Salida: Evaluación de los resultados de la minería de datos en lo que concierne a criterios de éxito de negocio**. Resumir los resultados de evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final en cuanto si el proyecto ya encuentra los objetivos iniciales de negocio.

c) **Modelos aprobados.** Después de la evaluación de modelos en lo que concierne a criterios de éxito de negocio, los modelos generados que encuentran los criterios seleccionados son los modelos aprobados.

2.5.2.5.2 Proceso de revisión

a) Tarea: Revisar el proceso. En este punto, los modelos resultantes pasan a ser satisfactorios y a satisfacer las necesidades de negocio. Ahora es apropiado hacer una revisión más cuidadosa de los compromisos de la minería de datos para determinar si hay cualquier factor importante o tarea que de algún modo ha sido pasada por alto. Esta revisión también cubre cuestiones de calidad;por ejemplo, ¿Construimos correctamente el modelo? ¿Usamos sólo los atributos que nos permitieron usar y que están disponibles para análisis futuros?

b) **Salida: Revisión de proceso.** Resumir la revisión de proceso y destacar las actividades que han sido omitidas y/o aquellas que deberían ser repetidas.

2.5.2.5.3 Determinación de los próximos pasos

a) Tarea: Determinar los próximos pasos. Según los resultados de la evaluación y la revisión de proceso, el equipo de proyecto decide cómo proceder. El equipo decide si hay que terminar este proyecto y tomar medidas sobre el desarrollo si es apropiado, tanto iniciar más iteraciones, o comenzar nuevos proyectos de minería de datos. Esta tarea incluye los análisis de recursos restantes y del presupuesto, que puede influir en las decisiones.

b) **Salida: Lista de posibles acciones.** Listar las acciones futuras potenciales, con los motivos a favor y en contra de cada opción.

c) **Decisión.** Describir la decisión en cuanto a cómo proceder, junto con el razonamiento.

2.5.2.6 Desarrollo

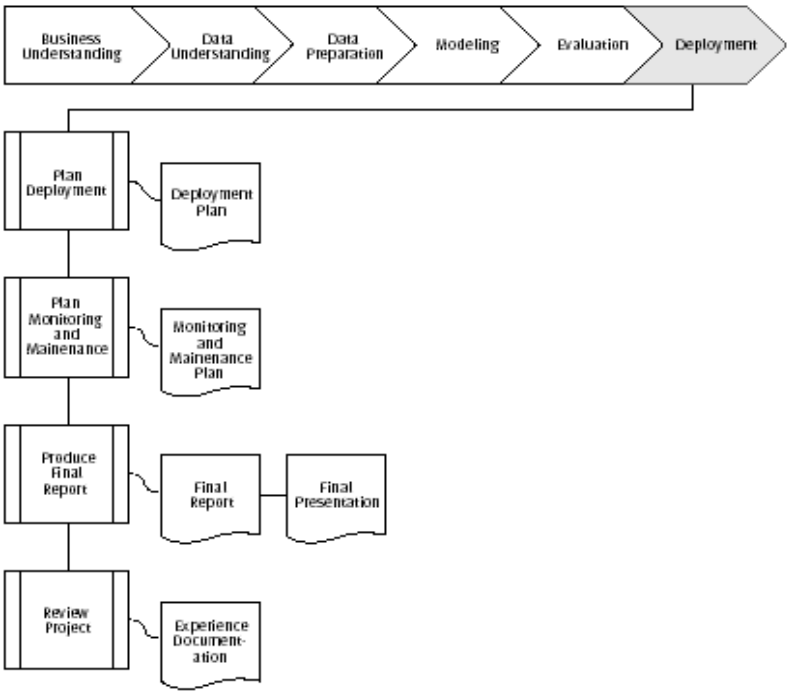


Figura Nro. 18, desarrollo.

2.5.2.6.1 Desarrollo del plan

a) Tarea: Desarrollar el plan. De acuerdo con el desarrollo de los resultados de minería de datos en el negocio, esta tarea toma los resultados de la evaluación y determina una estrategia para el desarrollo. Si un procedimiento general ha sido identificado para crear el/los modelo/s relevante/s, este procedimiento es documentado aquí para el desarrollo posterior.

b) **Salida: Desarrollo del plan.** Resumir la estrategia de desarrollo, incluyendo los pasos necesarios y como realizarlos.

2.5.2.6.2 Plan de supervisión y mantenimiento

a) Tarea: Planear la supervisión y el mantenimiento. La supervisión y el mantenimiento son cuestiones importantes si los resultados de minería de datos son parte del negocio cotidiano y de su ambiente. La preparación cuidadosa de una estrategia de mantenimiento ayuda evitar largos periodos innecesarios de uso incorrecto de resultados de minería de datos. Para

supervisar el desarrollo de los resultados de la minería de datos, el proyecto necesita un plan detallado de proceso de supervisión. Este plan tiene en cuenta el tipo específico de desarrollo.

b) **Salida: Supervisión y plan de mantenimiento.** Resumir la estrategia de supervisión y mantenimiento incluyendo los pasos necesarios y como realizarlos.

2.5.2.6.3 Informe definitivo de producto

a) Tarea: Producir el informe final. En el final del proyecto, el líder del proyecto y su equipo sobrescribe un informe final. Según el plan de desarrollo, este informe puede ser sólo un resumen del proyecto y sus experiencias (si estas aún no han sido documentadas como una actividad en curso) o esto puede ser una presentación final y comprensiva de los resultados de minería de datos.

b) **Salidas: Informe definitivo.** En el final del proyecto, el líder del proyecto y su equipo sobrescribe un informe final. Según el plan de desarrollo, este informe puede ser sólo un resumen del proyecto y sus experiencias (si estas aún no han sido documentadas como una actividad en curso) o esto puede ser una presentación final y comprensiva de los resultados de minería de datos.

c) **Presentación final.** También a menudo habrá una reunión en la conclusión del proyecto en el que los resultados son presentados verbalmente al cliente.

2.5.2.6.4 Revisión del proyecto

a) Tarea: Revisar el proyecto. Evaluar lo que fue correcto y lo que se equivocó, lo que fue bien hecho y lo que necesita para ser mejorado.

b) **Salida: Documentación de la experiencia.** Resumir las experiencias importantes ganadas durante el proyecto. Por ejemplo, trampas, accesos engañosos, o las insinuaciones para seleccionar las mejores técnicas de minería de datos en situaciones similares podrían ser la parte de esta documentación. En proyectos ideales, la documentación de la experiencia también cubre cualquier informe que ha sido escrito por miembros individuales del proyecto durante las fases del proyecto y sus tareas.

CAPÍTULO III

MÉTODO DE LA INVESTIGACIÓN

En este capítulo se detalla los pasos de la investigación que se realizaron para alcanzar al objetivo propuesto. La presente metodología cumple con cada una de las fases del modelo Crisp-DM, así mismo se detalla el plan de proyecto para su cumplimiento.

3.1 Tipo de investigación

La investigación que se está realizando es de dos tipos:

Tecnológica. Ya que consiste en la aplicación de los modelos y herramientas actuales a la solución del objeto de estudio.

Descriptivo. Se llegará a conocer la situación, identificación de estado económico en las facultades de la (UPeU) a través de la descripción exacta de las actividades, identificando la situación financieros de los indicadores Académica.

3.2 Diseño de la investigación

Luego de elegir la metodología Crisp-DM para la construcción, se decidió seguir la siguiente estrategia para la ejecución del proyecto la cual consiste en la realización de los siguientes seis fases: Comprensión del negocio, Comprensión de Datos, Preparación de datos, Modelado, Evaluación y Despliegue. Estos procesos se presentan en la figura 19 para la implantación de soluciones de inteligencia de negocios.

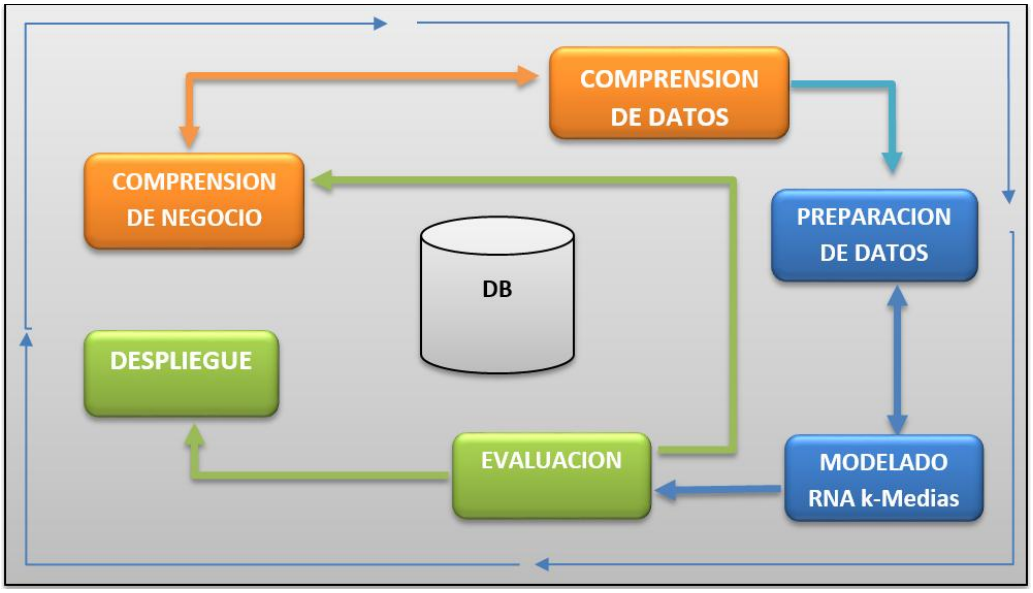


Figura 19, seis fases del diseño de investigación.

3.2.1 Comprensión del negocio.

En esta fase se obtiene conocimientos sobre las actividades del negocio del cliente, su estructura organizacional, ambiente interno, relaciones e interacciones y riesgos así como se muestra en la tabla 3.

Tabla Nro. 3, comprensión del negocio.

Comprensión del negocio	
ítem	**DESCRIPCION**
Contexto	Registro de la información requerido para el inicio del proyecto.
Objetivos del negocio	Análisis del negocio lo que el cliente realmente quiere lograr.
Inventario de recursos	Listar los recursos disponibles para el proyecto, incluyendo el personal (expertos de negocio, expertos de datos, soportes técnicos, expertos en minería de datos).
Riesgos y contingencias	Listar los riesgos o los acontecimientos que podrían retrasar el proyecto o hacer que ello falle
Terminología	compilamos un glosario de terminología relevante al proyecto
Costos	Construcción de un análisis de costo-beneficio para el proyecto
Objetivos y minería de datos	Describir las salidas intencionadas del proyecto que permiten el logro de los objetivos de negocio
Plan de proyecto	Listar las etapas a ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas, y dependencias.
Evaluación inicial	Evaluación inicial de herramientas y técnicas debería ser realizada. Aquí, por ejemplo, usted selecciona una herramienta de minería de datos que soporte varios métodos para las distintas etapas del proceso.

3.2.2 Se recolecta los datos para inicial el análisis del negocio

Tarea: Recolectar datos iniciales. Se adquiere en el proyecto los datos (o el acceso a los datos) listados en los recursos del proyecto. Esta colección inicial incluye carga de datos, si es necesario para la comprensión de los datos.

Salida: Informe de colección de datos inicial. Se lista el conjunto de datos adquiridos, juntos con sus posiciones, los métodos usados para adquirirlos, y algunos de los problemas encontrados. Luego se registra los problemas encontrados y algunas de las resoluciones alcanzadas. Esto ayudará la réplica (observación) futura de este proyecto o con la ejecución de proyectos similares futuros [25].

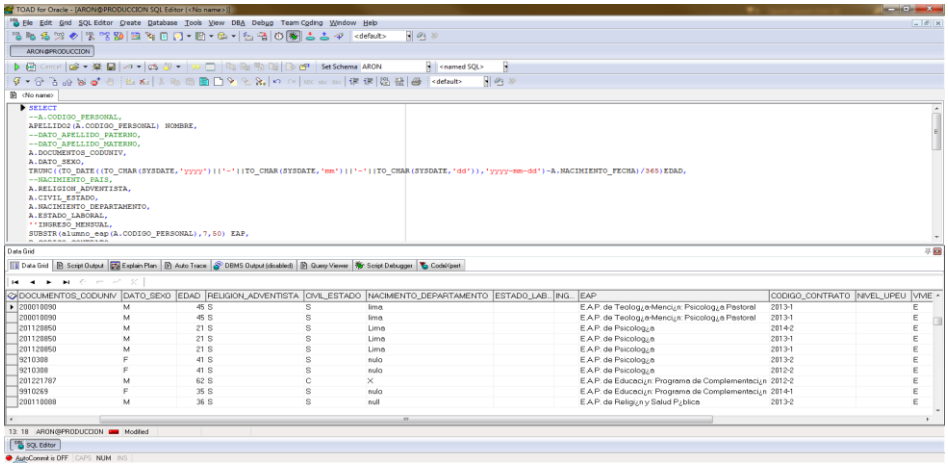


Figura 20, análisis inicial del negocio.

3.2.3 Preparación de datos.

c) **Salida: Conjunto de datos.** Este es el conjunto (o conjuntos) producido por la fase de preparación de datos, que será usada para modelar o para el trabajo principal de análisis del proyecto.

d) **Descripción del conjunto de datos.** Aquí se describe el conjunto de datos (o conjuntos) que será usado para el modelado y el trabajo principal de análisis del proyecto.

c) **Tarea: Selección de datos.** Decidir qué datos serán usados para el análisis. Los criterios incluyen la importancia de los objetivos de la minería de datos, la calidad, y las restricciones técnicas así como límites sobre el volumen de datos o los tipos de datos. Se nota que la selección de datos cubre la selección de atributos (columnas) así como la selección de registros (filas) en una tabla [26].

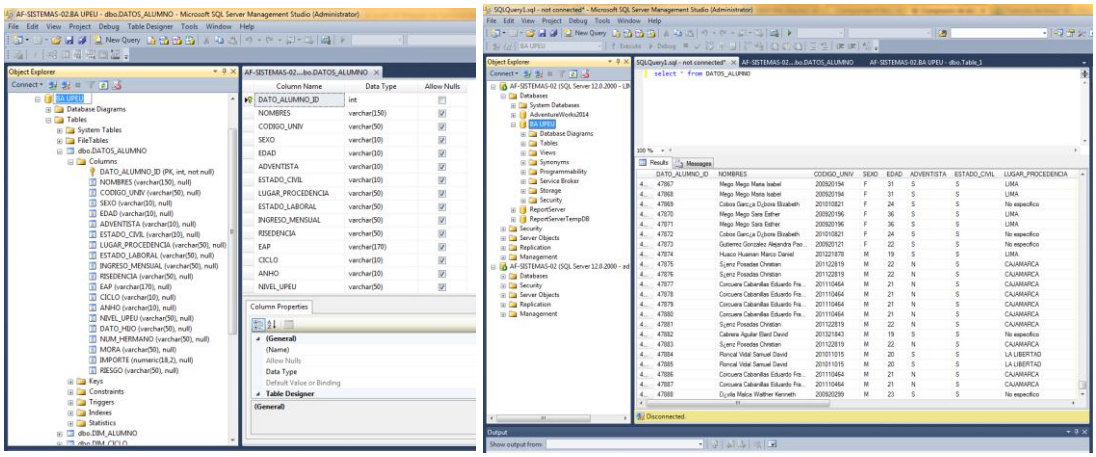


Figura 21, preparación de datos.

3.2.4 Modelado.

Tarea: Escoger la técnica de modelado. Esta tarea se refiere a la técnica de modelado específico, por ejemplo, un árbol decisión construido con Analysis Services, o la generación de red neuronales Back-Propagación. Si múltiples técnicas son aplicadas, se realizan esta tarea separadamente para cada técnica [25].

El proceso de agrupamiento K-Medias, inicialmente, se determina el número de grupos K y se asume el centroide o centro de grupos K. para determinar los centroides hay dos alternativas: la primera es tomar de forma aleatoria K objetos como centroides iniciales y la segunda es tomar los primeros K objetos en secuencia. Luego el algoritmo ejecuta los siguientes tres pasos hasta que alcance el criterio de convergencia; es decir, los objetos no se muevan del grupo [28]. Primero se determina los centroides iniciales de acuerdo con el número de clúster esperado, segundo se determina la distancia de cada objeto con relación a los centroides y la tercera se agrupan los objetos con base en la distancia mínima así como se muestra en la figura 22.

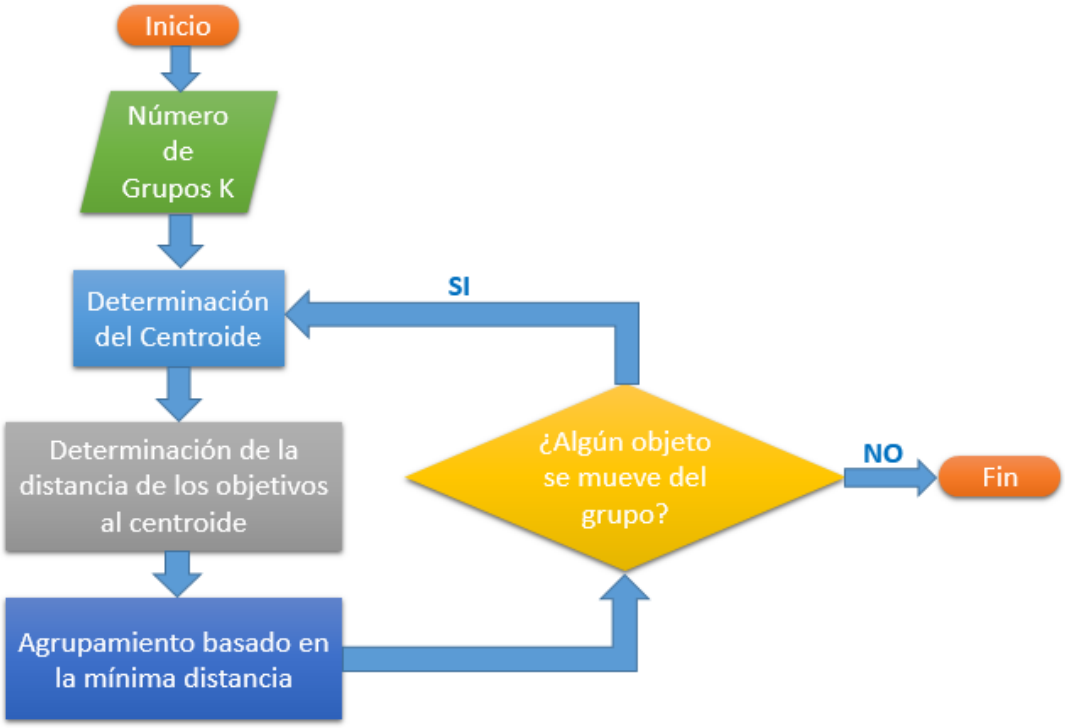


Figura Nro. 22, proceso de agrupacion de K-Medias.

Presunciones del modelado. Muchas técnicas de modelado hacen presunciones específicas sobre los datos; por ejemplo, todos los atributos tengan distribuciones uniformes, no encontrar valores no permitidos, el atributo de clase debe ser simbólico, etc. Registrar cualquiera de tales presunciones hechas [26].

3.2.5 Evaluación

Tarea: Evaluar los resultados. Los pasos de la evaluación anterior trata con factores como la exactitud y la generalidad del modelo. En este paso se evalúa el grado al que el modelo responde, los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el que este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten.

CAPÍTULO IV

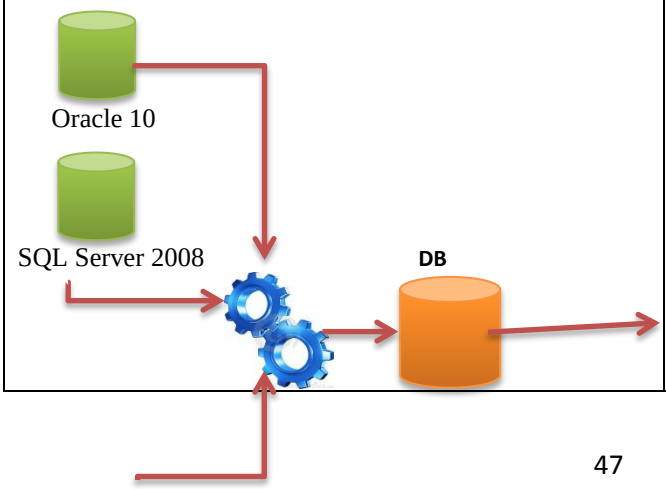
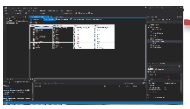
CONSTRUCCIÓN

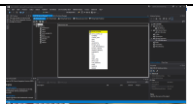
El capítulo presenta el proceso de construcción de BA. Estos son diseñados y desarrollados en cada uno de los pasos para la construcción de BA. Durante este capítulo hemos desarrollado en detalle cómo se modela los procesos utilizando la herramienta de Microsoft. De esta construcción de BA, se obtiene el resultado final para los usuarios finales, el cual permite para las tomas de decisiones de los gerentes del negocio.

4.1 Elaboración de BA para gerencia financiera de la UPeU.

A continuación se presenta la tabla de solución de BA (Business Analytics), muestra el diseño de solución de BA en tres fases: fase de definición, fase de diseño, Modelos y decisión. La cual comprende desde la extracción de la base de datos operacional y fuentes de formato xls. ETL, Creación de un nuevo DB (BA UPEU), tipos de información, muestra de información hasta obtener modelos según las características similares, así como se muestra en la tabla 3 diseños de solución de BA [29].

Tabla Nro. 4 – Diseño de solución de BA (Business Analytics).

Solución de BA (Business Analytics)		
Fuentes de Datos/fase de definición	**Fase de diseño**	**Modelo/ Decisión**
		

Encuesta	**ETL**		

4.1.1 Análisis de datos.

En la figura 23 se muestra el análisis de datos para nuevo base de datos (DATAWAREHOUSE), donde la información será exportada de la data operacional a través de ETL y que servirá como base para los diseños de modelos.

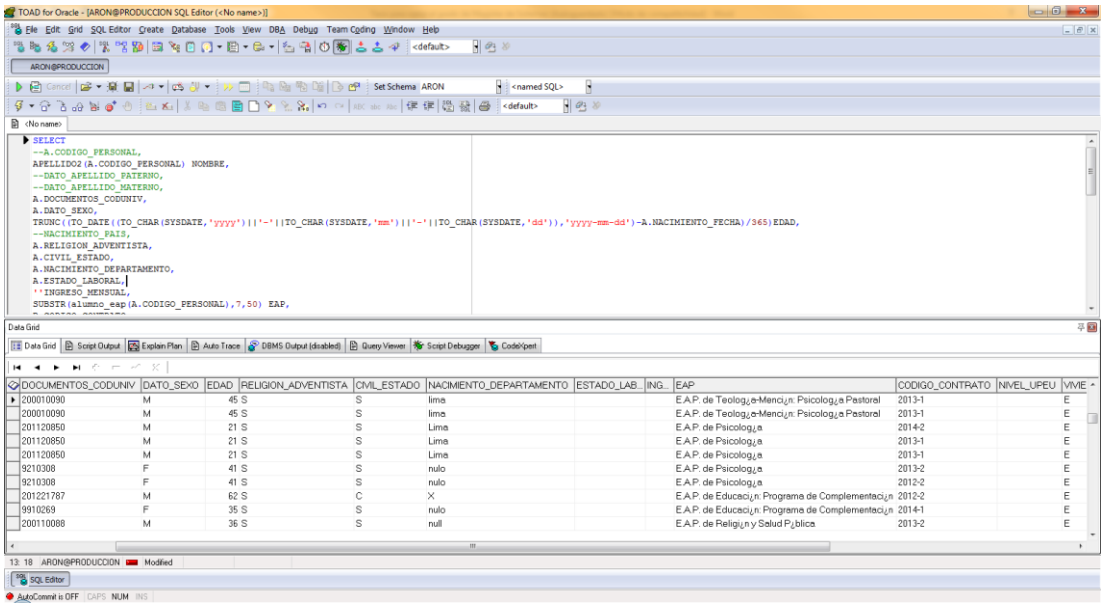


Figura 23 – Análisis de datos para la base de datos (DATAWAREHOUSE).

4.1.2 ETL (Extract Transform and Load).

La manipulación de la data operacional (Controladoria), es realizando ETL (Extract Transform and Load), donde es extraída la información a un Data Warehouse (BA UPEU), de esta forma no se cambia la data en la misma base de datos y la otra parte de datos es extraída de formato xls, para ETL utilizamos la herramienta de integración (Microsoft).

Figura 24 y 25 muestra la construcción de ETL para la carga de Datos a (DATAWAREHOUSE).

Figura 24 – Análisis de datos general.

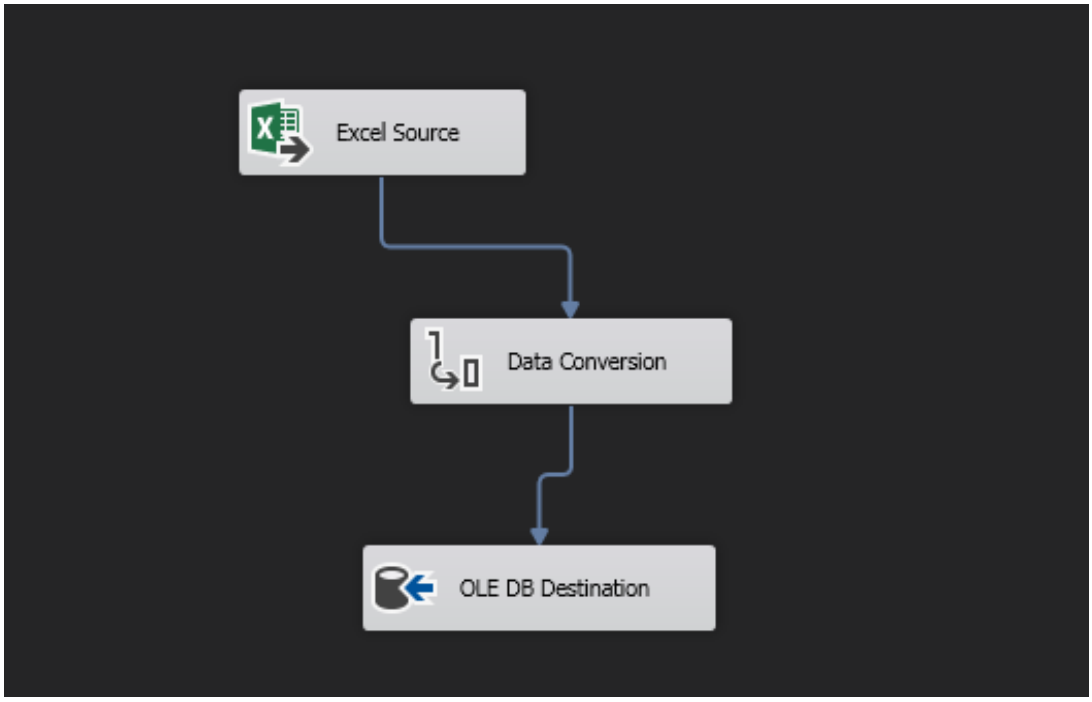


Figura 25, análisis de datos para la base de datos.

Después de análisis de datos y diseño en la figura 27 se prepara para la nueva DataWarehouse (BA UPEU), para carga de datos a través de ETL, esta base de datos es consolidada con la información requerida para la explotación de datos.

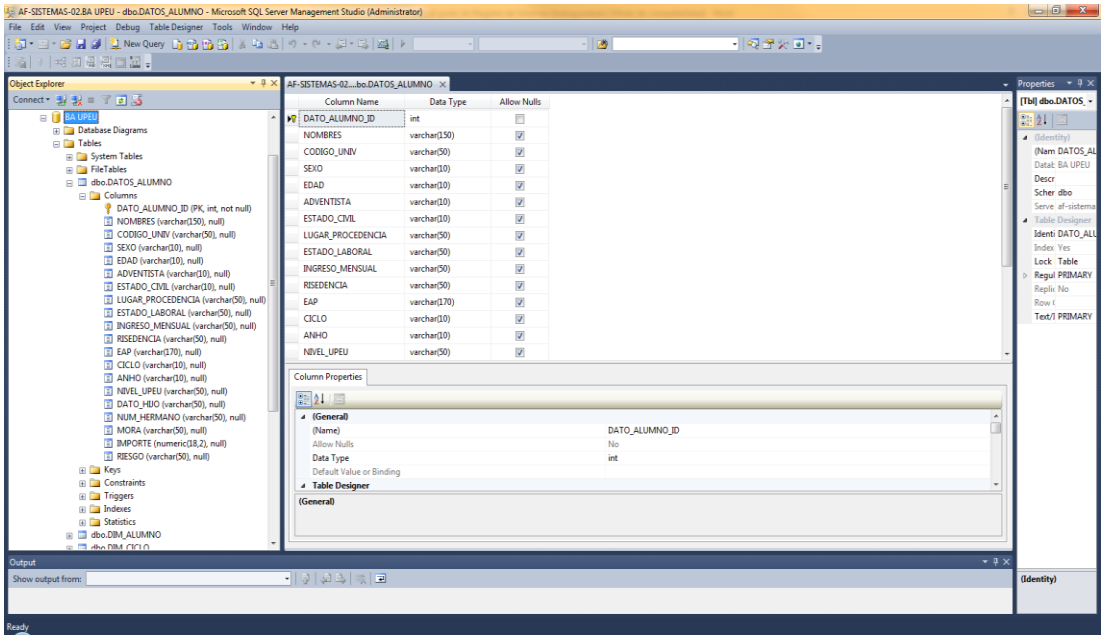


Figura 26, preparación de datawarehouse (BA UPEU).

En la figura 28 se procede a ejecutar el ETL para la carga de datos a una nueva base de datos (BA UPEU).

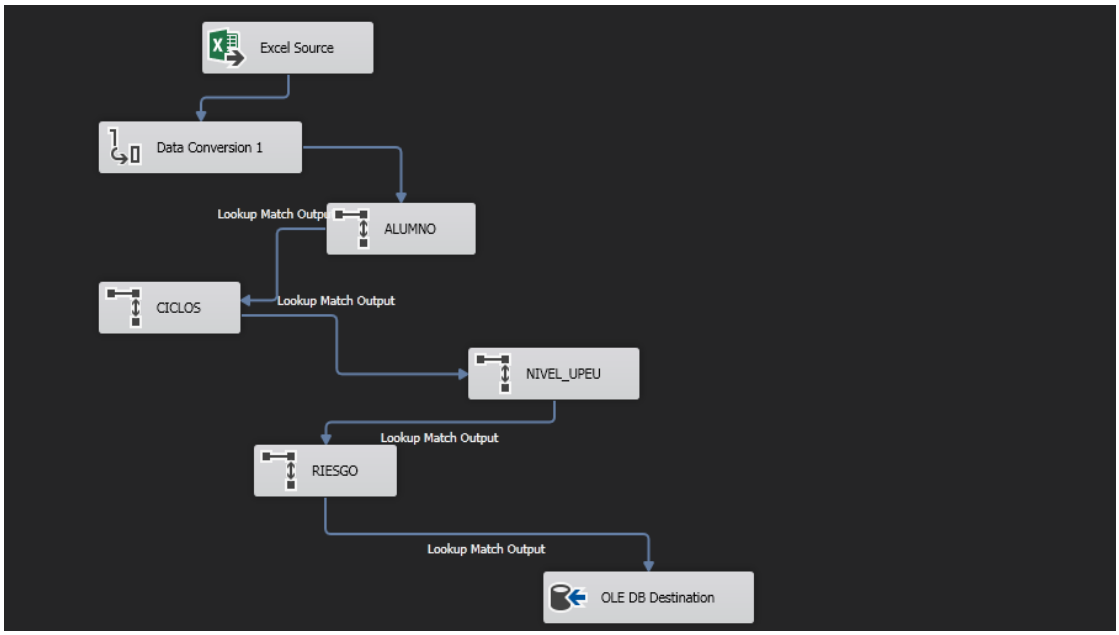


Figura 27, ETL para data warehouse (BA UPEU).

4.1.3 Fase de diseño de modelos.

Como el primero paso en modelado, se selecciona la técnica de modelado inicial actual según la herramienta a utilizar, para procesar nuestro modelo se hace la conexión a la Data Warehouse (BA UPEU), así como se muestra en la figura 28.

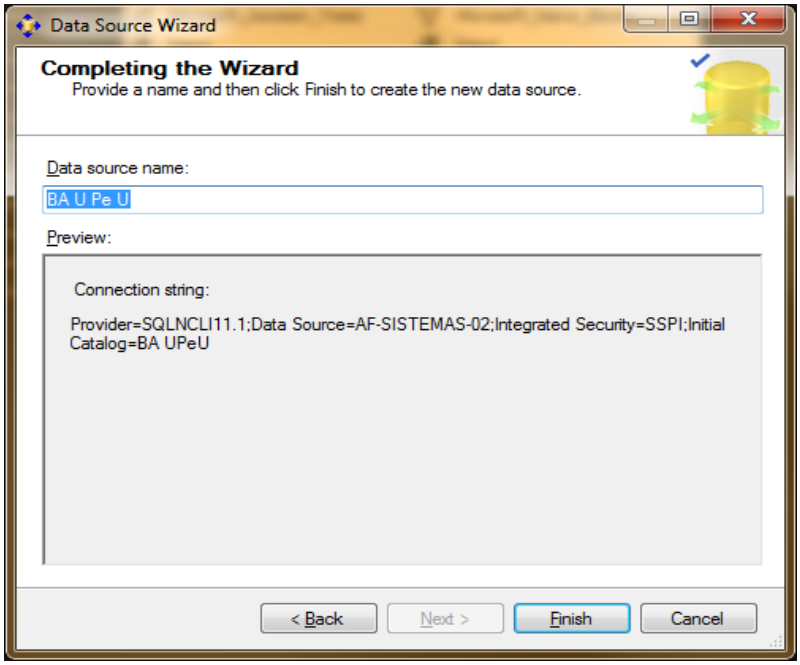


Figura 28, conexión a base de datos.

En la figura 29, se muestra la selección de tabla donde está almacenada los datos del alumno.

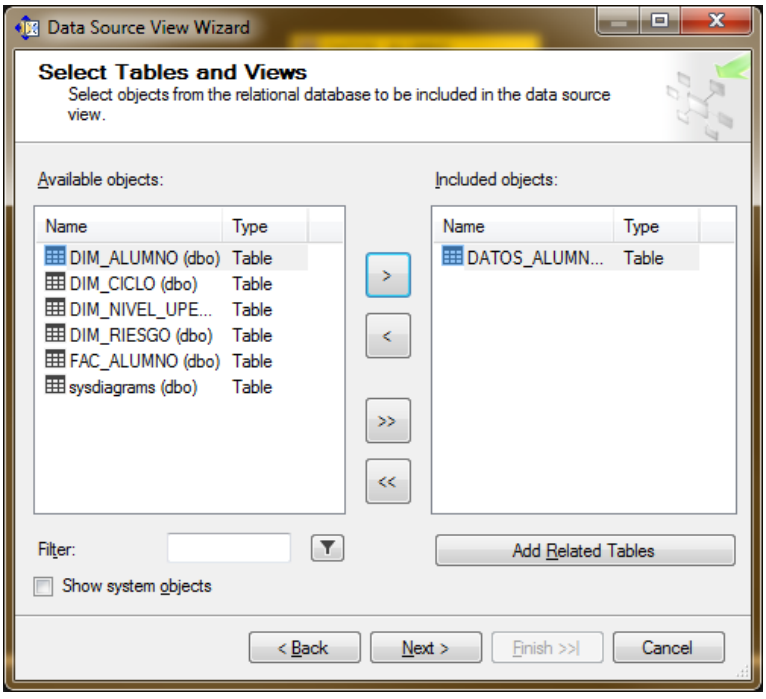


Figura 29, selección de tabla para el modelo.

Se selecciona la técnica de modelado inicial actual, según la herramienta a utilizar, procesamos nuestro modelo de agrupamiento con el algoritmo de k-medias. En la figura 30 se muestra la selección de la técnica en la herramienta.

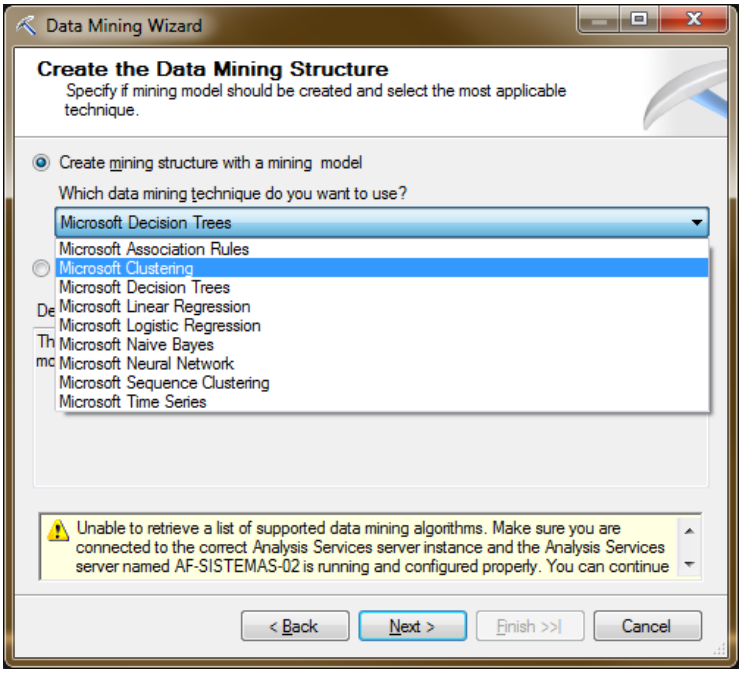


Figura 30, selección del algoritmo para el modelo.

Se procede a generar el diseño de prueba, ya que antes de construir un modelo, es necesario definir un procedimiento para probar la calidad del modelo y la validez. El diseño de prueba específica que el conjunto de datos debería ser separado en el entrenamiento y en el conjunto de prueba. El modelo está construido sobre el conjunto de entrenamiento y su calidad estimada sobre el conjunto de prueba. En la figura 31 se muestra los datos de aprendizaje y el conjunto de pruebas del modelo.

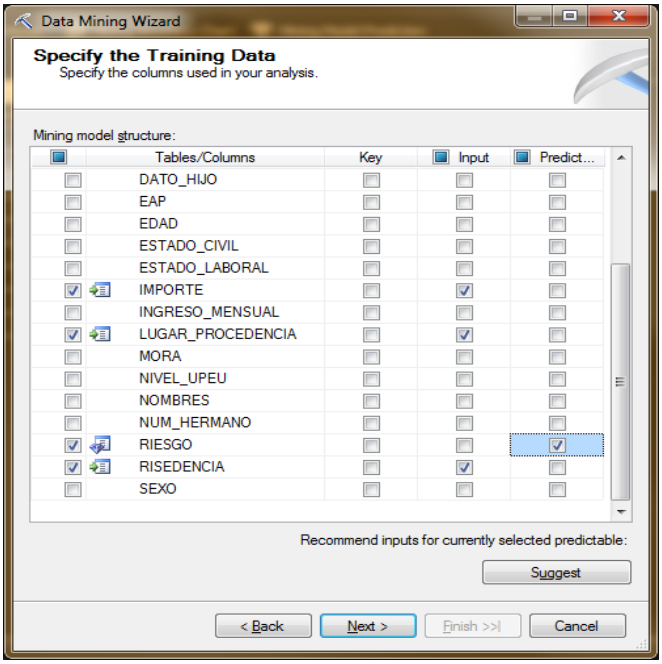


Figura 31, muestra la creación del conjunto de datos y de pruebas que se utiliza para aprendizaje del modelo.

En la figura 32, se muestra la sugerencia de datos seleccionados

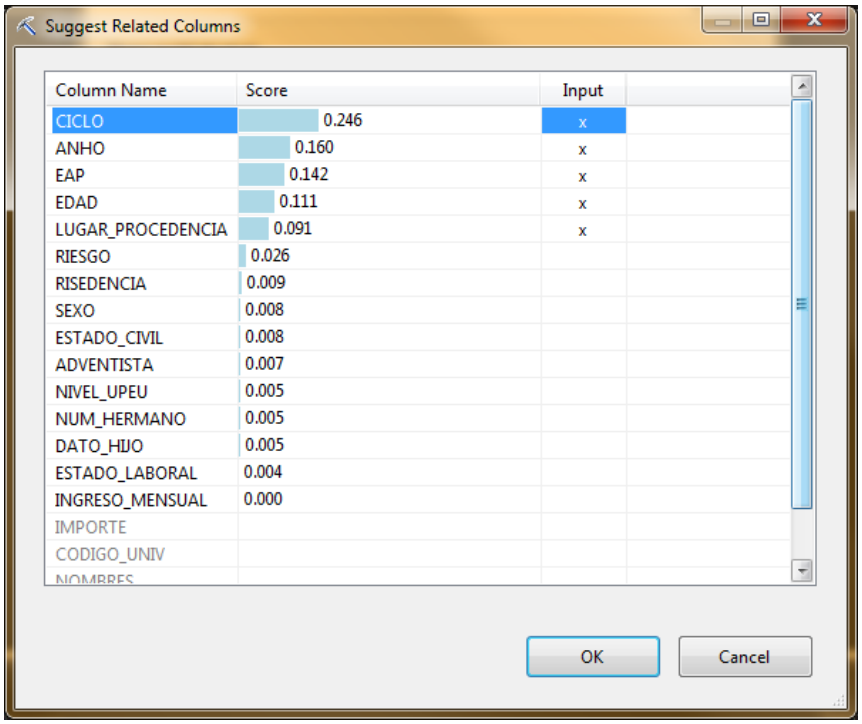


Figura 32, sugerencia de datos seleccionados.

Se procede a la construcción del modelo, para ello se utiliza la herramienta de modelado sobre el conjunto de datos, hay a menudo un gran número de parámetros que pueden ser ajustados. En la figura 33, se muestra la el procesamiento de los clústeres de este modelo.

Structure	DATOS ALUMNO	ARBOLES	NBAYES
	Microsoft_Clustering	Microsoft_Decision_Trees	Microsoft_Naive_Bayes
ADVENTISTA	Input	Input	Input
ANHO	Input	Input	Input
CICLO	Input	Input	Input
DATO ALUMNO ID	Key	Key	Key
EDAD	Input	Input	Input
ESTADO CIVIL	Input	Input	Input
IMPORTE	Input	Input	Input
MORA	Input	Input	Input
RIESGO	PredictOnly	PredictOnly	PredictOnly

Figura 33 – Procesamiento de modelos.

CAPÍTULO V

VALIDACIÓN Y RESULTADOS

El presenta capítulo tiene el objetivo de presentar la validación de cada uno de los resultados solicitados para el área financiera académica de la (UPeU). De igual forma se detalla los reportes dinámicos de la información para los indicadores propuestos.

Para los reportes de modelos se utilizó la herramienta Analysis Services, donde se realiza la selección de algoritmos, para realizar el reporte de modelo.

Se procedió a identificar los clústeres que tienen la mayor cantidad de alumnos morosos y les asignamos sus nombres correspondientes, para hacer un contraste entre los alumnos que no deben y aquellos que sí. En la figura 34, se muestra los clústeres generados por la herramienta.

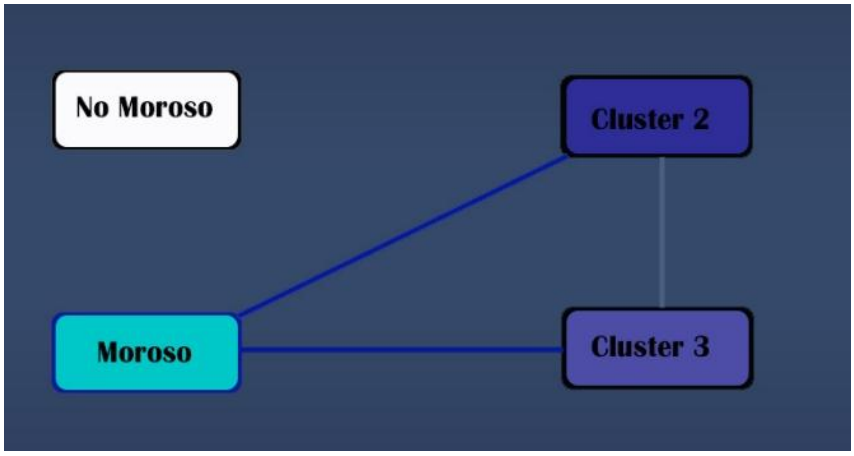


Figura Nro. 34, agrupación de modelos.

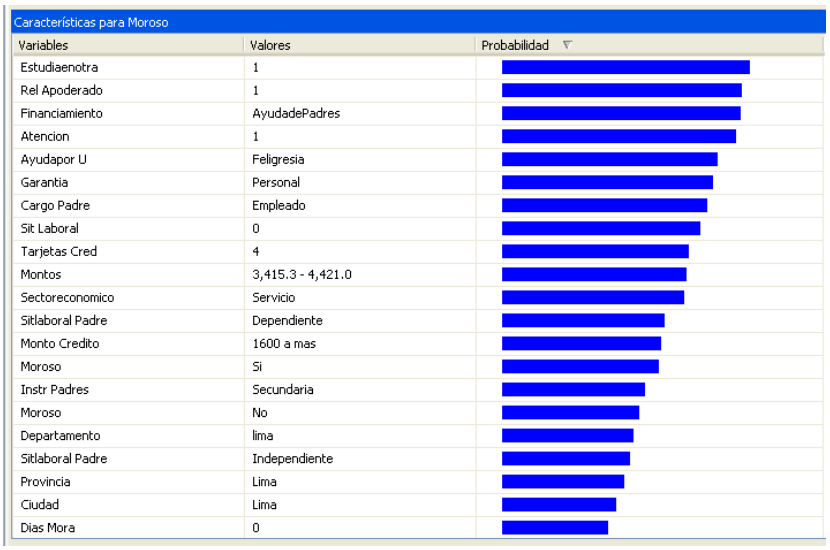
Una vez identificado los clústeres se procedió a analizar las características más resaltantes de los alumnos morosos. De acuerdo con la información recopilada se puede observar lo siguiente:

5.1 Análisis de los atributos

5.1.1 Clúster moroso

Un clúster es la agrupación de casos de alumnos que tienen características similares. En la figura 35 se puede observar los atributos más importantes presentes mediante probabilidades y se pueden ir de siguiente modo:

a) La probabilidad de que el alumno sea moroso si es que estudia en otra institución es 84.211%.
b) La probabilidad de que el alumno sea moroso y el responsables del pago sea su padre es de 81.115%
c) Una variable que favorece a la morosidad es si el alumno trabaja, por lo tanto la probabilidad de que el alumno no trabaje influya en la morosidad es de 67.492%
d) La probabilidad de que sea moroso dado que el cargo laboral de su apoderado es Empleado es de 69.659%



Características para Moroso

Variables	Valores	Probabilidad
Estudiaenotra	1	
Rel Apoderado	1	
Financiamiento	AyudadePadres	
Atencion	1	
Ayudapor U	Feligresia	
Garantia	Personal	
Cargo Padre	Empleado	
Sit Laboral	0	
Tarjetas Cred	4	
Montos	3,415.3 - 4,421.0	
Sectoreconomico	Servicio	
Sitlaboral Padre	Dependiente	
Monto Credito	1600 a mas	
Moroso	Si	
Instr Padres	Secundaria	
Moroso	No	
Departamento	lima	
Sitlaboral Padre	Independiente	
Provincia	Lima	
Ciudad	Lima	
Dias Mora	0	

Figura Nro. 35, ficha características del clúster.

5.1.2 Clúster no moroso

En este clúster se detallan las características que influyen en la no morosidad del alumno, como observaremos en la Figura 36, la ficha de características del clúster se puede analizar y conocer para la toma de desiciones, según se detallan a continuación:

a) La probabilidad de que el alumno no sea moroso dado que trabaja es de 100.00%
b) La probabilidad de que el alumno no sea moroso dado que su apoderado tiene instrucción secundaria es de 100.00%
c) La probabilidad de que el alumno no sea moroso dado que su apoderado tiene un ingreso bruto entre S/ 1010.5 y 1432 es de 100.00%
d) La probabilidad de que el alumno no sea moroso dado que tiene apoyo de sus padres es de 70.00%

Ahora se analiza las características de ambos tipos de alumnos identificando las diferencias más resaltantes:

a) Favorece la situación laboral de jubilado a padre del alumno no moroso.
b) Que el alumno trabaje favorece al clúster no moroso.

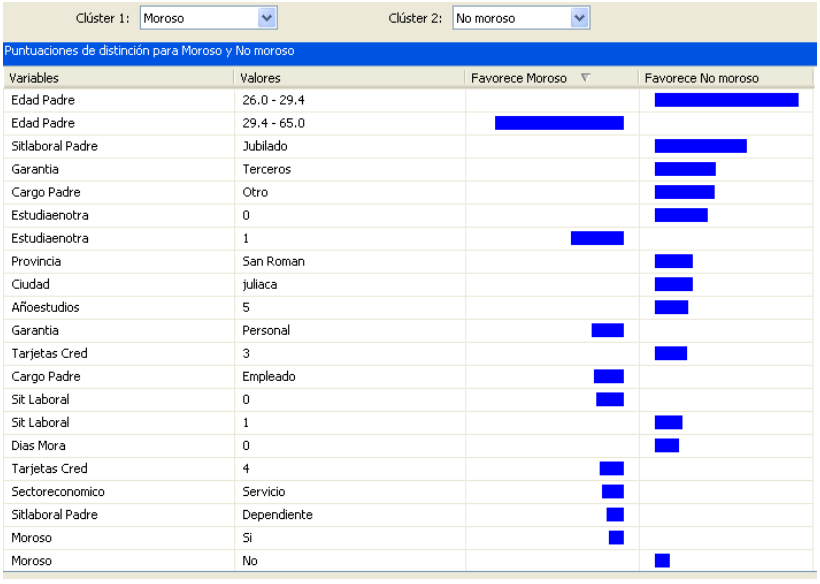


Figura Nro. 36, Ficha distinción del clúster.

Analisis de morosidad consolidada. El ánalisis se muestra en la siguiente Figura 37, por sedes, filiales y centros de aplicación.

MOROSIDAD Setiembre del 2015									
#	MOROSIDAD	INGRESO 2014	SALDO 2014	INGRESO 2015	SALDO 2015	% 2014	% 2015	% Var.	
CONSOLIDADO									
1.	UNIVERSIDAD SEDE	28775143,46	3,277,448.10	31,345,961.80	3,462,057.88	11.00	11.00	.00	
2.	CENTRO DE APLICACION IMPRENTA UNION	7389592,54	-876,032.72	5,592,610.83	256,111.63	-12.00	5.00	141.67	
3.	COLEGIO UNION	2529616,49	457,870.57	3,367,816.47	561,838.21	18.00	17.00	5.56	
4.	FILIAL JULIACA	10013400,96	1,304,135.14	11,828,877.58	1,490,933.63	13.00	13.00	.00	
5.	FILIAL TARAPOTO	6079589,61	761,349.79	6,163,276.11	605,154.79	13.00	10.00	23.08	
6.	INSTITUTO DE IDIOMAS	1266520,84	166,621.53	1,809,462.20	375,457.70	13.00	21.00	61.54	
7.	CONSERVATORIO DE MUSICA	257287	32,522.00	331,727.10	35,207.00	13.00	11.00	15.38	
8.	PROESAD	4254617,7	657,281.67	5,142,154.52	778,569.55	15.00	15.00	.00	
9.	POSTGRADO	2858295,1	558,619.34	3,722,480.99	458,496.45	20.00	12.00	40.00	

Figura Nro. 37, análisis de morosidad consolidad.

Analisis de morosidad por facultades. El análisis se muestra en la siguiente Figura 38, por facultades de sede Lima, así como se observa el la figura 38.

MOROSIDAD Setiembre del 2015						
#	MOROSIDAD	SALDO	TOTAL ALMUNOS	INDICE	% MORO	IND.
FACULTADES						
1.	FACULTAD CC EMPRESARIALES	662713,9	763	868,56	10.51	
2.	FACULTAD DE INGENIERIA	923944,43	923	1001,02	10.53	
3.	FACULTAD DE CIENCIAS DE LA SALUD	1101792,02	949	1161	10.76	
4.	FACULTAD DE HUMANAS Y EDUCACION	296758,5	351	845,47	11.26	
5.	FACULTAD DE TEOLOGIA	474006,45	376	1260,66	14.05	
6.	GESTION PROESAD	778569,55	1634	476,48	15.23	

Figura Nro. 38, análisis de morosidad por facultades.

Analisis de morosidad por Escuela. El análisis se muestra en la siguiente Figura 39, por escuelas de la sede Lima, asi como se observa el la figura.

MOROSIDAD Setiembre del 2015						
#	MOROSIDAD	SALDO	TOTAL ALMUNOS	INDICE	% MORO	IND.
FACULTAD CC EMPRESARIALES						
1.	UN CTP Informatica Empresarial	29,451.02	46	640,24	12.00	
2.	UN CTP Secretariado Ejecutivo	44,280.90	95	466,11	14.00	
3.	UN EAP Administracion	310,828.82	345	900,95	10.00	
4.	UN EAP Contabilidad	278,153.16	226	1230,77	11.00	
FACULTAD DE INGENIERIA						
5.	UN EAP Arquitectura	299,517.66	226	1325,3	12.00	
6.	UN EAP Ingenieria Ambiental	280,410.30	282	994,36	11.00	
7.	UN EAP Ingenieria Civil	153,311.47	65	2358,64	9.00	
8.	UN EAP Ingenieria de Alimentos	71,528.35	88	812,82	11.00	
9.	UN EAP Ingenieria de Sistemas	119,176.65	173	688,88	9.00	

Figura Nro. 39, análisis de morosidad por escuela.

Análisis de morosidad comparativos. El siguiente análisis comparativo es por escuelas de la sede Lima, así como se observa el la Figura 40.

MOROSIDAD

Setiembre del 2015

#	MOROSIDAD	% 2014	% 2015	% Var.
	FACULTAD CC EMPRESARIALES			
1.	UN EAP Contabilidad	13.00	11.00	15.38
2.	UN EAP Administracion	11.00	10.00	9.09
3.	UN CTP Informatica Empresarial	14.00	12.00	14.29
4.	UN CTP Secretariado Ejecutivo	15.00	14.00	6.67
	FACULTAD DE INGENIERIA			
5.	UN EAP Ingenieria de Sistemas	10.00	9.00	10.00
6.	UN EAP Ingenieria de Alimentos	10.00	11.00	10.00
7.	UN EAP Ingenieria Ambiental	10.00	11.00	10.00
8.	UN EAP Arquitectura	10.00	12.00	20.00
9.	UN EAP Ingenieria Civil	9.00	9.00	.00

Figura Nro. 40, análisis de morosidad comparativo.

5.2 Análisis de los resultados obtenidos estadísticos.

Estadístico de T de prueba con respecto al pago de cuotas según las fechas establecidas en el contrato.

Para la selección de la muestra se ha utilizado los números aleatorios generando mediante el SPSS, para conocer el resultado se ha considerado una muestra de 130, para conocer la morosidad del alumno.

Sub – hipótesis 1.

H_0: Se desea comprobar la puntuación media del coeficiente de la morosidad del grupo de alumnos que paga las cuotas según la fecha establecida en su contrato (antes de los 25 días).

H_1: Se desea comprobar la puntuación media del coeficiente de la morosidad del grupo de alumnos que no paga las cuotas según la fecha establecida en su contrato (después de los 25 días).

Esto es:

H_0: $\mu \leq 25$

H_1: $\mu > 25$

Nivel de significación $\boldsymbol{\alpha}$ = 0.05

Estadístico de prueba es t-student $\mathbf{T_{n-1}}$ = 4.98

P ó Sig = error teórico establecido (0.05 ó 0.01)

Cálculo usando SPSS.

Tabla Nro. 5, estadística de T de prueba para la comprobación de la morosidad del alumno.

Prueba estadístico

	N	Media	Desviación Típica	Error tip. de la media
Morosidad (dias)	130	38.32	30.503	2.675

Prueba de test de muestras

	Test Value = 25					
	t	Df	Sig. (2-tailed)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Condición del alumno (Morosidad)	4.98	129	,000	13.323	8.03	18.62

Criterio de decisión

Si p < α entonces se rechaza la hipótesis nula

Se Rechaza la hipótesis $\mathbf{H_0}$. Prueba unilateral, p=0.000/2 =0; el valor de t = 4.98 tendrá un p = 1-0 = 1 > 0.05, entonces se rechaza la hipótesis nula $\mathbf{H_0}$. y aceptamos la hipótesis alterna.

Entonces se acepta la hipótesis alterna $\mathbf{H_1}$, se comprueba que la mayoría de los alumnos paga su cuota después de la fecha indicada en su contrato.

Estadístico T de prueba para comprobar la morosidad con respecto a ingreso bruto mensual

Para selección de la muestra se ha utilizado los números aleatorios generando mediante el SPSS, para conocer el resultado se ha considerado una muestra de 130, para conocer la morosidad del alumno.

Sub – hipótesis 2.

H_0: Si el alumno trabaja tendrá un ingreso bruto mensualmente, entonces habrá menor probabilidad que el alumno sea moroso.

H_1: Si el alumno no trabaja no tendrá un ingreso bruto mensualmente, entonces habrá mayor probabilidad que el alumno sea moroso.

H_0: $\mu t \leq$ um

H_1: $\mu t >$ um

Nivel de significación $\boldsymbol{\alpha} = 0.05$

P ó Sig = error teórico establecido (0.05 ó 0.01)

Cálculo usando SPSS.

Tabla Nro. 6, Grupo estadístico de situación laboral.

Situación laboral		N	Media	Desviación estándar	Error tip. de la media
Ingreso Bruto	Si trabaja	28	1885.71	1545,774	292.124
	No trabaja	102	1533.50	565,157	55.959

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	T	Gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia Superior	Inferior
Ingreso Bruto	Se han asumido varianzas iguales	16.36	,000	1,899	128	,596	144,743	271,758	-398,671	688,157
	No se han asumido varianzas iguales			1,184	29,007	,740	144,743	427,606	-778,392	1067,878

Decisión: como p = 0.011 < 0.05 entonces se rechaza la hipótesis nula $\mathbf{H_0}$.

En conclusión, el grupo de los alumnos que no trabajan presentó mayor probabilidad de que sean morosos que el grupo de los alumnos que trabajan. Si los alumnos que no trabajan μt es mayor la morosidad *um* en 144.743 puntos.

5.2.1 Identificación del contexto de la población.

Tabla Nro. 7, Recibe ayuda por parte de universidad.

	Frecuencia	porcentaje	Porcentaje válidos	Porcentaje acumulado
Beca de estudios	25	19,2	19,2	19,2
Beca de feligresía	85	65,4	65,4	84,6
Ningún ayuda	20	15,4	15,4	100,0
Total	130	100,0	100,0	

Se puede observar del grafico que el 65.38% de los 85 encuestados reciben beca de feligresía, los 19.23% de los 25 encuestados recibe becas de estudio (1/4, 1/2, 1 becas) de parte de la universidad y 15.38% de los 20 encuestados no recibe ninguna ayuda por parte de la universidad.

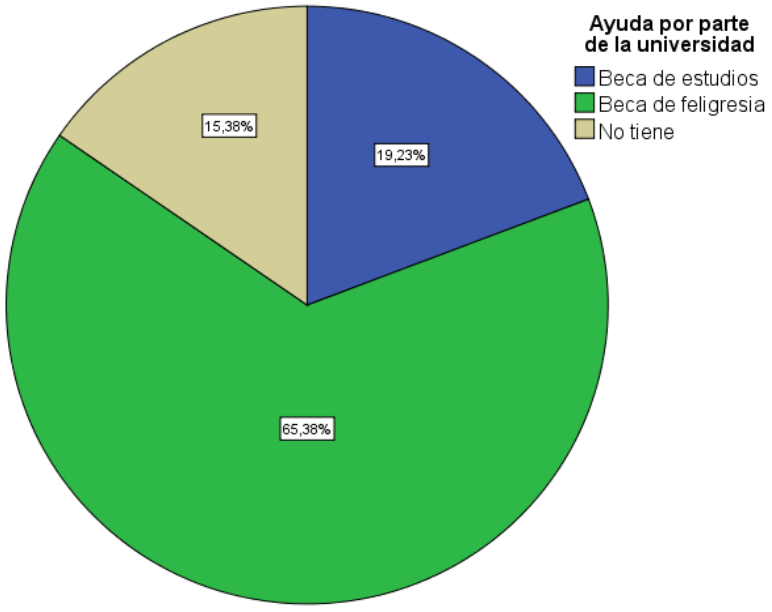


Figura Nro. 41, Resultados obtenidos ayuda por parte de Universidad.

Se concluye que la mayoría de los entrevistados recibe ayuda de la universidad, tanto con beca de feligresía y becas de estudios (1/4, 1/2, 1 beca) y muy poco de los encuestados no recibe ninguna ayuda.

Tabla Nro. 8, hermanos estudiando en la universidad.

	Frecuencia	porcentaje	Porcentaje válidos	Porcentaje acumulado
uno	23	17,7	17,7	17,7
dos	19	14,6	14,6	32,3
tres	7	5,4	5,4	37,7
No tiene	81	62,3	62,3	100,0
Total	130	100,0	100,0	

Fuente: cuestionario propuesto.

Se observa que el 63.31% de los 81 encuestado no tiene ningún hermano estudiando en la UPeU, el 17.69% de los 23 encuetados tiene un hermano estudiando en la universidad, el 14.62% de los 19 encuestados tiene 2 hermanos estudiando y los 5.38% de los 7 encuestados tiene tres hermanos estudiando en la universidad.

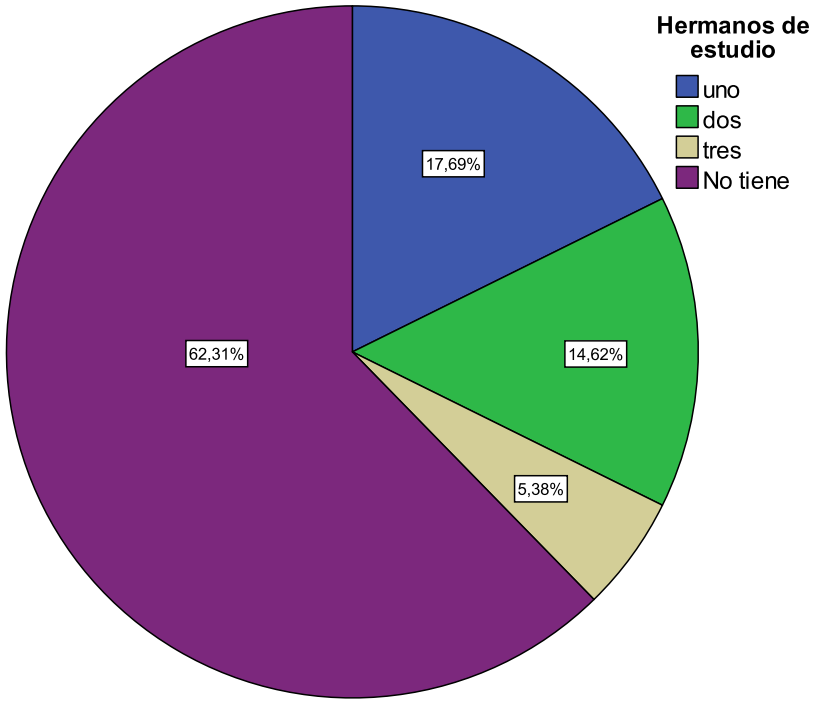


Figura Nro. 42, Resultados obtenidos hermanos de estudio.

En conclusión, los estudiantes quienes tienen más de un hermano estudiando en la universidad, es muy probable que tenga deficiencias de pago de sus cuotas de mensualidad.

Tabla Nro. 9, Financiamiento de estudio.

	Frecuencia	porcentaje	Porcentaje válidos	Porcentaje acumulado
Ayuda de padres	91	70,0	70,0	70,0
Beca de ayuda	11	8,5	8,5	78,5
Auto sostenimiento	19	14,6	14,6	93,1
Beca ley	9	6,9	6,9	100,0
Total	130	100,0	100,0	

Se muestra que el 70.00% de los 91 encuestados recibe ayuda de sus padres, el 14.62% de los 19 alumnos encuetados se auto sostiene, 8.46% de los 11 encuestados recibe ayuda de becas y el 6.92% de los 9 encuestados recibe Beca ley.

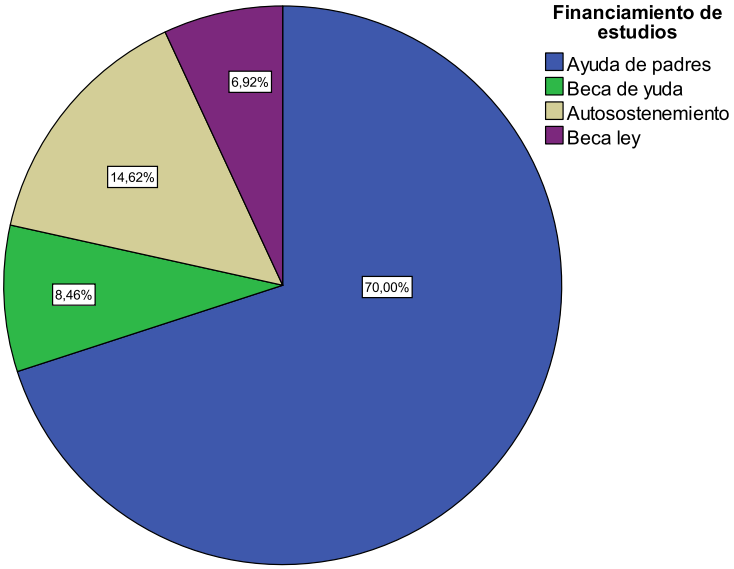


Figura Nro. 43, Resultados obtenidos hermanos de estudio.

Se concluye: los alumnos que se auto sostienen hay una mayor probabilidad que sea moroso, porque no recibe ayuda de parte de sus apoderados.

CAPÍTULO VI

CONCLUSIONES Y RECOMENDACIONES

1. Conclusiones

a) Después de haber realizado el trabajo de investigación, utilizando la herramienta de BA, se ha podido experimentar y conocer la segmentación de clústeres de BA, que ha facilitado el trabajo desde la fase de definición, la fase de diseño, hasta la fase de explotación de la información de modelos para la toma de decisiones. De este modo, se ha logrado el objetivo propuesto inicialmente, y se ha logrado una correcta integración con cada una de las herramientas utilizadas en el proyecto de investigación.

b) Con la creación de un modelo de clúster, y la elaboración de BA (Business Analytics) se ha podido lograr mejorar en las tomas de decisiones, facilitando así el manejo dinámico de los reportes, las operaciones de navegación que son bastante flexibles, cuyos usuarios finales interactúan y la herramienta muestra la información requerida para la toma de decisiones, para conocer el resultado se ha considerado una muestra de 130, para conocer la morosidad del alumno la fecha de pago antes y después de 25 de cada mes obteniendo un resultado de clúster moroso para después de 38.32 días.

   Para conocer el resultado se ha considerado una muestra de 130, para conocer la morosidad del alumno si trabaja o no trabaja obteniendo un resultado de clúster moroso tiene un ingreso bruto mensual 1533.50 y de clúster no moroso tiene un ingreso de 1885.71

c) Se logró identificar los clústeres con características similares de los alumnos de la Universidad Peruana Unión, los cuales fueron capturados de la base de datos relacional, acorde con la necesidad de información a predecir.

2. Recomendaciones

a) Utilizar las herramientas de Analysis Service BI (Analytics Service), y la metodología Crisp-DM para garantizar la calidad de modelo de clúster y del trabajo en otras investigaciones similares.
b) Se recomienda que al momento de manipular la información de la data histórica para enviar a una nueva data, sacar el backup, porque una mala manipulación puede alterar la información de la base de datos, asimismo por cuestiones de seguridad.
c) La solución abarca el área financiera de la Universidad, se recomienda continuar con la misma metodología y herramienta aplicada a otras áreas de la (UPeU).
d) La puesta en marcha del modelo de Clúster pueda seguir un proceso continuo para desarrollar la iniciativa y, posteriormente, ejecutar y de ese modo monitorear en toda la corporación.

REFERENCIAS BIBLIOGRÁFICAS

[1] J. R. Santos, “Sistemas de Soporte a la Decisión (Business Intelligence) para las Pymes de Collado Villalba,” pp. 1–18, 2009.

[2] A. Bhattacharya and R. K. De, “Divisive correlation clustering algorithm (DCCA) for grouping of genes: Detecting varying patterns in expression profiles,” *Bioinformatics*, vol. 24, no. 11, pp. 1359–1366, 2008.

[3] Consorcio Cluster Development, “Elaboración de un mapeo de clústers en el Perú. Consultoría solicitada por el Consejo Nacional de la Competitividad.,” pp. 1–342, 2013.

[4] H. Crc, *DATA CLUSTERING Algorithms and Apllications*. 2014.

[5] A. Luis, A. Romero, and T. Calonge, “Capítulo 1.- Redes Neuronales y Reconocimiento de Patrones.,” pp. 1–11.

[6] A. Alzate and E. Giraldo, “Clasificación de Arritmias utilizando ANFIS, Redes Neuronales y Agrupamiento Substractivo,” *Sci. Tech.*, vol. 12, no. 31, pp. 19–22, 2006.

[7] J. J. Montaño, “Redes Neuronales Artificiales aplicadas al Análisis de Datos,” *Network*, p. 275, 2002.

[8] D. J. Matich, “Redes Neuronales: Conceptos Básicos y Aplicaciones.,” *Historia Santiago.*, p. 55, 2001.

[9] J. D. Meisel and L. K. Prado, “UN ALGORITMO GENÉTICO HÍBRIDO Y UN ENFRIAMIENTO DE PROGRAMACIÓN DE PEDIDOS JOB SHOP J osé D avid M eisel * L iliana K atherine P rado **,” *Rev. EIA*, vol. 13, pp. 39–51, 2010.

[10] V. O. S. Rodallegas Ramos Erika, Torres González Areli ,. Gaona Couto Beatriz B, Gastelloú Hernández Erick, Lezama Morale Rafael As, “Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos,” in *Recursos Digitales*, M. E. P. / J. M. D. / D. O. Villegas, Ed. UNIVERSIDAD TECNOLÓGICA

METROPOLITANA, 2010.

[11] F. Izaurieta and C. Saavedra, “Redes Neuronales Artificiales,” *Charlas Fis.*, pp. 1–15, 1999.

[12] B. J. Haas and M. C. Zody, “Advancing RNA-Seq analysis.,” *Nature biotechnology*, vol. 28, no. 5. pp. 421–423, 2010.

[13] K. S. Kosik, “Circles reshape the RNA world,” *Nature*, vol. 495, pp. 4–6, 2013.

[14] R. X. and D. C. Wunsch, *Clustering*. Wiley, IEEE Press series on computational Intelligence, 2009.

[15] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *J. Intell. Inf. Syst.*, vol. 17, no. 2–3, pp. 107–145, 2001.

[16] H. Jiawei and M. Kamber, *Data mining: concepts and techniques*. 2001.

[17] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[18] J. Pérez, M. F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas, and A. Mexicano, “Mejora al algoritmo de agrupamiento K-means mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer,” in *Liver- 2do Taller Latino Iberoamericano de Investigacion de Operaciones “la IO aplicada a la solución de problemas regionales”.*, 2007, pp. 1–7.

[19] M. X. Dueñas-Reye, “Minería de datos espaciales en búsqueda de la verdadera informaciùn. (Spanish),” *Ing. y Univ.*, vol. 13, pp. 137–156, 2009.

[20] H. Zha, X. He, C. Ding, and H. Simon, “Spectral Relaxation for K-means Clustering,” *MIT Press*, pp. 1057–1064, 2001.

[21] P. Chapman, J. Clinton, and R. Kerber, “CRISP-DM 1.0: Cross Industry Standard Process for Data Mining,” 2000.

[22] S. Moro and R. M. S. Laureano, “Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology,” in *European Simulation and Modelling Conference*, 2011, no. Figure 1, pp. 117–121.

[23] A. Azevedo and M. F. Santos, “KDD, SEMMA and CRISP-DM: a parallel overview,” *IADIS Eur. Conf. Data Min.*, pp. 182–185, 2008.

[24] B. Carneiro da Rocha and R. Timoteo de Sousa Junior, “Identifying Bank Frauds Using CRISP-DM and Decision Trees,” *International Journal of Computer Science and Information Technology*, vol. 2, no. 5. pp. 162–169, 2010.

[25] C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining,” *J. Data Warehous.*, vol. 5, no. 4, pp. 13–22, 2000.

[26] R. Wirth, “CRISP-DM : Towards a Standard Process Model for Data Mining,” *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, pp. 29–39, 2000.

[27] J. C. M. Zapata and N. Gil, “Incorporation of both pre-conceptual schemas and goal diagrams in CRISP-DM,” in *2011 6th Colombian Computing Congress, CCC 2011*, 2011.

[28] K. Krishna and M. N. Murty, “Genetic K-means algorithm,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 29, no. 3, pp. 433–439, 1999.

[29] G. H. N. Laursen and J. Thorlund, *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting*. 2010.

ANEXOS

Anexo 1 – Presupuesto de proyecto de investigación.

Recursos de Personal	**Meses**	**Sueldo Mensual (S/.)**	**Sub Total (S/.)**
Analista de datos	2	S/. 2500.00	S/. 5000.00
Modelamiento de datos	2	S/. 2000.00	S/. 4000.00
Inscripción del proyecto			S/. 100.00
Asesor del Proyecto			S/. 800.00
Sustentación			S/. 700.00
Total gastos en personal			**S/. 10600.00**
Recursos de Software	**Cantidad**	**Precio Unitario (S/.)**	**Sub Total (S/.)**
SQL Server 2008	1	S/. 5.00	S/. 5.00
Pentaho	**1**	S/. 5.00	S/. 5.00
Toad	1	S/. 5.00	S/. 5.00
Total gastos en software			**S/. 15.00**
Recursos de Hardware	**Cantidad**	**Precio Unitario (S/.)**	**Sub Total (S/.)**
Computadoras	2	S/. 400.00	S/. 400.00
Impresor	1	S/. 270.00	S/. 270.00
USB	1	S/. 50.00	S/. 50.00
DVDs y CDs	6	S/. 1.50	S/. 9.00
Cartucho de la impresora.	1	S/. 70.00	S/. 70.00
½ millar Papel A4	500	S/. 12.00	S/. 12.00
Impresión	370	S/. 37.00	S/. 37.00
Total gastos en hardware			**S/. 848.00**
Gastos Administrativos	**Mes**	**Precio Unitario (S/.)**	**Sub Total (S/.)**
Internet (moden)	6	S/. 99.00	S/. 594.00
Transporte	6		S/. 200.00
Imprevistos	6		S/. 50.00
Gastos Indirectos	6		S/. 50.00
Total gastos administrativos			**S/. 894.00**
Gran total			**S/. 12357.00**

Anexo 2 - Consulta datos alumnos

```
SELECT
--A.CODIGO_PERSONAL,
APELLIDO2(A.CODIGO_PERSONAL) NOMBRE,
--DATO_APELLIDO_PATERNO,
--DATO_APELLIDO_MATERNO,
A.DOCUMENTOS_CODUNIV,
A.DATO_SEXO,
TRUNC((TO_DATE((TO_CHAR(SYSDATE,'yyyy')||'-
'||TO_CHAR(SYSDATE,'mm')||'-'||TO_CHAR(SYSDATE,'dd')),'yyyy-mm-dd')-
A.NACIMIENTO_FECHA)/365)EDAD,
--NACIMIENTO_PAIS,
A.RELIGION_ADVENTISTA,
A.CIVIL_ESTADO,
A.NACIMIENTO_DEPARTAMENTO,
A.ESTADO_LABORAL,
''INGRESO_MENSUAL,
SUBSTR(alumno_eap(A.CODIGO_PERSONAL),7,50) EAP,
B.CODIGO_CONTRATO,
''NIVEL_UPEU,
B.VIVIENDA_TIPO,
A.DATO_HIJOS,
B.NUM_HERMANO,
B.MORA,
C.IMPORTE,
''ANHO,
''RIESGO
--RELIGION_ALUMNO
FROM DATOS_PERSONALES A, ALUMNO_CONTRATO B, UPEU_SALDO_INI C
WHERE A.CODIGO_PERSONAL = B.CODIGO_PERSONAL
AND A.CODIGO_PERSONAL = C.ID_PERSONAL
AND B.CODIGO_CONTRATO IN (
'2012-1',
'2012-2',
'2013-1',
'2013-2',
'2014-1',
'2014-2',
'2015-1',
'2015-2'
)
AND DOCUMENTOS_CODUNIV IS NOT NULL
--AND B.AREA_ID = '2'
GROUP BY
A.CODIGO_PERSONAL, A.DOCUMENTOS_CODUNIV,
A.DATO_SEXO, A.NACIMIENTO_FECHA,
A.RELIGION_ADVENTISTA, A.ESTADO_LABORAL,
A.NACIMIENTO_DEPARTAMENTO, A.CIVIL_ESTADO,
B.VIVIENDA_TIPO, A.DATO_HIJOS,
B.NUM_HERMANO, B.CODIGO_CONTRATO,
B.MORA, C.IMPORTE
```

Anexo 3 - Consulta de Morosidad Consolidado Total.

```
SELECT
        UNI_ID,
        (CASE UNI_ID
          WHEN '0' THEN 'CONSOLIDADO'
          WHEN '10000000' THEN 'LIMA'
          WHEN '20000000' THEN 'PRODUCTOS UNION'
          WHEN '30000000' THEN 'IMPRENTA UNION'
          WHEN '40000000' THEN 'COLEGIO UNION'
          WHEN '50000000' THEN 'JULIACA'
          WHEN '60000000' THEN 'TARAPOTO'
        END) DEPTO,
        (CASE CO
          WHEN 1 THEN 'CAPITAL OPERATIVO'
          WHEN 2 THEN 'SOSTENIMIENTO PROPIO'
        END) NOMBRE,
        TIE_ANHO,
        TIE_MES,
        abs(CAPITAL) as CAPITAL,
        (
         CASE
              WHEN CO = 1 AND UNI_ID = 0 THEN 1
              WHEN CO = 2 AND UNI_ID = 0 THEN 2
              WHEN CO = 1 AND UNI_ID = 10000000 THEN 3
              WHEN CO = 2 AND UNI_ID = 10000000 THEN 4
              WHEN CO = 1 AND UNI_ID = 20000000 THEN 5
              WHEN CO = 2 AND UNI_ID = 20000000 THEN 6
              WHEN CO = 1 AND UNI_ID = 30000000 THEN 7
              WHEN CO = 2 AND UNI_ID = 30000000 THEN 8
              WHEN CO = 1 AND UNI_ID = 40000000 THEN 9
              WHEN CO = 2 AND UNI_ID = 40000000 THEN 91
              WHEN CO = 1 AND UNI_ID = 50000000 THEN 92
              WHEN CO = 2 AND UNI_ID = 50000000 THEN 93
              WHEN CO = 1 AND UNI_ID = 60000000 THEN 94
              WHEN CO = 2 AND UNI_ID = 60000000 THEN 95
              END
        ) ORDEN
    FROM (
        SELECT
            1 AS CO,
            A.UNI_ID,A.IND_ID,B.TIE_ANHO,B.TIE_MES, ROUND((A.FAC_IMPORTE/
```

```
                (
                 SELECT X.FAC_IMPORTE
                 FROM BI_FAC_INDICADOR X
                 WHERE X.UNI_ID = A.UNI_ID
                 AND X.TIE_ID = B.TIE_ID
                 AND X.IND_ID = '000077'
                )*100),3) AS CAPITAL
        FROM BI_FAC_INDICADOR A, BI_TIEMPO B
        WHERE A.TIE_ID = B.TIE_ID
        AND B.TIE_ANHO IN (2011,2012)
        AND B.TIE_MES <= ".09."
        AND A.IND_ID = '000030'
        UNION ALL
        SELECT
       2                AS                SP,A.UNI_ID,A.IND_ID,B.TIE_ANHO,B.TIE_MES,
ROUND((A.FAC_IMPORTE/
                (
                 SELECT X.FAC_IMPORTE
                 FROM BI_FAC_INDICADOR X
                 WHERE X.UNI_ID = A.UNI_ID
                 AND X.TIE_ID = B.TIE_ID
                 AND X.IND_ID = '000039'
                )*100),3) AS CAPITAL
        FROM BI_FAC_INDICADOR A, BI_TIEMPO B
        WHERE A.TIE_ID = B.TIE_ID
        AND B.TIE_ANHO IN (2011,2012)
        AND B.TIE_MES <= ".09."
        AND A.IND_ID = '000038'
        ORDER BY UNI_ID,TIE_MES
 )
 ORDER BY ORDEN,TIE_ANHO,TIE_MES ";
```

Anexo 4 - Consulta Morosidad por facultad.

```
SELECT
        'FACULTADES' FAC_NOMBRE,
        (SELECT X.UNI_NOMBRE FROM sim.bi_unidad_estrategica X WHERE
SUBSTR(X.UNI_ID,1,4) = FAC_ID AND X.UNI_NIVEL_GESTION = '2')
UNI_NOMBRE,
        FAC_ID,
        UIN_IMPORTE,
        UIN_IMPORTE_ACUM,
        TO_CHAR(UIN_PORCENTAJE*100,'999,999,999,999,999.99')
UIN_PORCENTAJE,
        TOTAL,
        ROUND((TO_NUMBER(UIN_IMPORTE_ACUM)/TO_NUMBER(TOTAL)),2)
INDICE,
 (CASE WHEN UIN_PORCENTAJE <= 18 AND
ROUND((TO_NUMBER(UIN_IMPORTE_ACUM)/TO_NUMBER(TOTAL)),2) <= 1000
THEN 1
        WHEN UIN_PORCENTAJE <= 18 AND
ROUND((TO_NUMBER(UIN_IMPORTE_ACUM)/TO_NUMBER(TOTAL)),2) > 1000
THEN 2
        WHEN UIN_PORCENTAJE BETWEEN 18 AND 21 AND
ROUND((TO_NUMBER(UIN_IMPORTE_ACUM)/TO_NUMBER(TOTAL)),2)
BETWEEN 1000 AND 2000 THEN 2
        WHEN UIN_PORCENTAJE > 21 AND
ROUND((TO_NUMBER(UIN_IMPORTE_ACUM)/TO_NUMBER(TOTAL)),2) > 2000
THEN 0
        END ) AS SEM_XX
 FROM (
          SELECT
                        SUBSTR(A.UNI_ID,1,4) AS FAC_ID,
                        SUM(B.UIN_IMPORTE) UIN_IMPORTE,
                        SUM(B.UIN_IMPORTE_ACUM) UIN_IMPORTE_ACUM,
                        (SUM(B.UIN_IMPORTE_ACUM)/SUM(B.UIN_IMPORTE))
UIN_PORCENTAJE,
SUM(SIM.FC_CW_TOTAL_ALUMNOS_EAP('".$mai_id."','".$nivel."','".$semestre."',A.U
NI_ID)) AS TOTAL
          FROM sim.bi_unidad_estrategica A, sim.bi_unidad_indicador B,sim.bi_tiempo C
          WHERE A.UNI_ID = B.UNI_ID
          AND B.tie_id = C.tie_id
          AND B.ind_id = '000037'
```

```
            AND A.uni_estado = '1'
            AND C.tie_anho = '".$anho."'
            AND C.tie_mes = '".$mes_id."'
            AND SUBSTR(A.UNI_ID,1,3) = '".$sede."'
            GROUP BY SUBSTR(A.UNI_ID,1,4)
)
ORDER BY FAC_ID, UNI_NOMBRE";
```

Anexo 5 - Consulta Morosidad por escuela.

```
select UNI_ID,UIN_IMPORTE,UIN_IMPORTE_ACUM,UIN_PORCENTAJE from (
        select
            aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-'||'".$anho."')
    UNI_ID,
            sum(importe) UIN_IMPORTE,
            sum(saldo) UIN_IMPORTE_ACUM,
            round(sum(saldo)/decode(sum(importe),0,1,sum(importe)),2)
    UIN_PORCENTAJE
        from (
        select
            id_personal,
            sum(abs(nvl(importe,0))) importe,
            0 saldo
            from aron.upeu_mov_doc
        where id_venta = '001-'||'".$anho."'
        and to_char(fecha,'MM') <= lpad('".$i."',2,0)
        and id_personal in (
        select
                id_personal
            from sim.bi_data
            where id_venta = '001-'||'".$anho."'
            and id_contrato like '".$anho."'||'".$semestre."'
            and tipo = 'R'
        )
        and importe <> 0
        and dc = 'D'
        and tipo = 'S'
        group by id_personal
        having sum(abs(nvl(importe,0))) <> 0
        union all
        Select
            id_personal,
            0 importe,
            decode(sign(sum(importe)),-1,0,sum(importe)) saldo
        from aron.upeu_sal_cli_mes
        where id_venta = '001-'||'".$anho."'
        and to_char(fecha,'MM') <= lpad('".$i."',2,0)
        and id_personal in (
```

```
                select
                        id_personal
                    from sim.bi_data
                    where id_venta = '001-'||'".$anho."'
                    and id_contrato like '".$anho."'||'".$semestre."'
                    and tipo = 'R' )
            group by id_personal
            having decode(sign(sum(importe)),-1,0,sum(importe)) <> 0
            )
            group by aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-
'||'".$anho."')
        )
        where UNI_ID not like '130103%'
        and UNI_ID like '130%'
        union
        select UNI_ID,UIN_IMPORTE,UIN_IMPORTE_ACUM,UIN_PORCENTAJE from (
        select
            aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-'||'".$anho."')
UNI_ID,
            sum(importe) UIN_IMPORTE,
            sum(saldo) UIN_IMPORTE_ACUM,
            round(sum(saldo)/decode(sum(importe),0,1,sum(importe)),2)
UIN_PORCENTAJE
            from (
            select
                id_personal,
                sum(abs(nvl(importe,0))) importe,
                0 saldo
                from aron.upeu_mov_doc
            where id_venta = '001-'||'".$anho."'
            and to_char(fecha,'MM') <= lpad('".$i."',2,0)
            and id_personal in (
            select
                    id_personal
                from sim.bi_data
                where id_venta = '001-'||'".$anho."'
                and tipo = 'P'
            )
            and importe <> 0
            and dc = 'D'
            and tipo = 'S'
```

```
group by id_personal
having sum(abs(nvl(importe,0))) <> 0
union all
Select
    id_personal,
    0 importe,
    decode(sign(sum(importe)),-1,0,sum(importe)) saldo
from aron.upeu_sal_cli_mes
where id_venta = '001-'||'".$anho."'
and to_char(fecha,'MM') <= lpad('".$i."',2,0)
and id_personal in (
    select
            id_personal
        from sim.bi_data
        where id_venta = '001-'||'".$anho."'
        and tipo = 'P' )
group by id_personal
having decode(sign(sum(importe)),-1,0,sum(importe)) <> 0
)
group by aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-'||'".$anho."')
)
where UNI_ID like '14%'
union
select UNI_ID,UIN_IMPORTE,UIN_IMPORTE_ACUM,UIN_PORCENTAJE from (
select
    aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-'||'".$anho."') UNI_ID,
    sum(importe) UIN_IMPORTE,
    sum(saldo) UIN_IMPORTE_ACUM,
    round(sum(saldo)/decode(sum(importe),0,1,sum(importe)),2) UIN_PORCENTAJE
from (
select
    id_personal,
        sum(abs(nvl(importe,0))) importe,
        0 saldo
    from aron.upeu_mov_doc
where id_venta = '001-'||'".$anho."'
and to_char(fecha,'MM') <= lpad('".$i."',2,0)
and id_personal in (
```

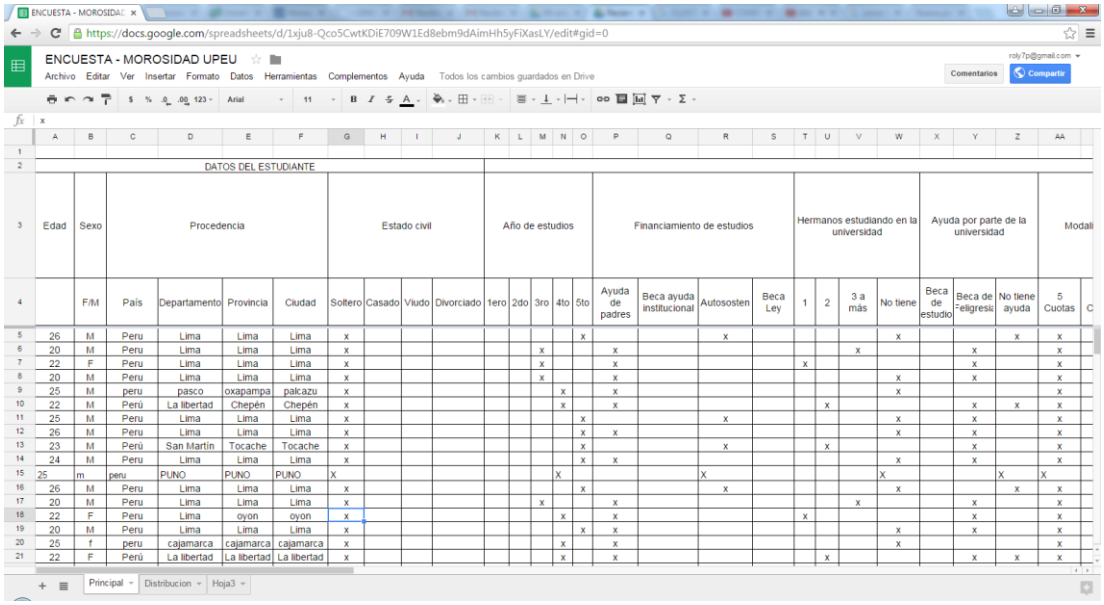
```
                select
                        id_personal
                from sim.bi_data
                where id_venta = '001-'||'".$anho."'
                and id_contrato like '".$anho."'||'".$semestre."'
                and tipo = 'A'
        )
        and importe <> 0
        and dc = 'D'
        and tipo = 'S'
        group by id_personal
        having sum(abs(nvl(importe,0))) <> 0
        union all
        Select
            id_personal,
            0 importe,
            decode(sign(sum(importe)),-1,0,sum(importe)) saldo
        from aron.upeu_sal_cli_mes
        where id_venta = '001-'||'".$anho."'
        and to_char(fecha,'MM') <= lpad('".$i."',2,0)
        and id_personal in (
            select
                    id_personal
                from sim.bi_data
                where id_venta = '001-'||'".$anho."'
                and id_contrato like '".$anho."'||'".$semestre."'
                and tipo = 'A' )
        group by id_personal
        having decode(sign(sum(importe)),-1,0,sum(importe)) <> 0
        )
        group by aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-
'||'".$anho."')
    )
    --where UNI_ID like '130103%'
    where UNI_ID like '130602%'
    union
    select UNI_ID,UIN_IMPORTE,UIN_IMPORTE_ACUM,UIN_PORCENTAJE from (
    select
        aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-'||'".$anho."')
UNI_ID,
        sum(importe) UIN_IMPORTE,
```

```
                sum(saldo) UIN_IMPORTE_ACUM,
                round(sum(saldo)/decode(sum(importe),0,1,sum(importe)),2)
UIN_PORCENTAJE
        from (
        Select
                id_personal,
                sum(abs(nvl(importe,0))) importe,
                0 saldo
            from aron.tara_mov_doc
            where id_venta = '001-'||'".$anho."'
            and to_char(fecha,'MM') <= lpad('".$i."',2,0)
            and id_personal in (
                select
                    id_personal
                from sim.bi_data
                where id_venta = '001-'||'".$anho."'
                and id_contrato like '".$anho."'||'".$semestre."'
                and tipo = 'T'
        )
            and importe <> 0
            and dc = 'D'
            and tipo = 'S'
            group by id_personal
            having sum(abs(nvl(importe,0))) <> 0
            union all
            Select
                id_personal,
                0 importe,
                decode(sign(sum(importe)),-1,0,sum(importe)) saldo
            from aron.tara_sal_cli_mes
            where id_venta = '001-'||'".$anho."'
            and to_char(fecha,'MM') <= lpad('".$i."',2,0)
            and id_personal in (
                select
                    id_personal
                from sim.bi_data
                where id_venta = '001-'||'".$anho."'
                and id_contrato like '".$anho."'||'".$semestre."'
                and tipo = 'T'
        )
            group by id_personal
```

```
                having decode(sign(sum(importe)),-1,0,sum(importe)) <> 0
            )
            group by aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-'||'".$anho."')
        )
        where UNI_ID like '630%'
        union
        select UNI_ID,SUM(UIN_IMPORTE) UIN_IMPORTE,SUM(UIN_IMPORTE_ACUM) UIN_IMPORTE_ACUM,SUM(UIN_PORCENTAJE) UIN_PORCENTAJE from (
        select
            aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-'||'".$anho."') UNI_ID,
            sum(importe) UIN_IMPORTE,
            sum(saldo) UIN_IMPORTE_ACUM,
            round(sum(saldo)/decode(sum(importe),0,1,sum(importe)),2) UIN_PORCENTAJE
        from (
        Select
                id_personal,
                sum(abs(nvl(importe,0))) importe,
                0 saldo
            from aron.chullu_mov_doc
            where id_venta = '001-'||'".$anho."'
            and to_char(fecha,'MM') <= lpad('".$i."',2,0)
            and id_personal in (
                select
                    id_personal
                from sim.bi_data
                where id_venta = '001-'||'".$anho."'
                and id_contrato like '".$anho."'||'".$semestre."'
                and tipo = 'J'
        )
            and importe <> 0
            and dc = 'D'
            group by id_personal
            having sum(abs(nvl(importe,0))) <> 0
            union all
            Select
                id_personal,
                0 importe,
```

```
                    decode(sign(sum(importe)),-1,0,sum(importe)) saldo
                from aron.chullu_mov_doc
                where id_venta = '001-'||'".$anho."'
                and to_char(fecha,'MM') <= lpad('".$i."',2,0)
                and id_personal in (
                    select
                        id_personal
                    from sim.bi_data
                    where id_venta = '001-'||'".$anho."'
                    and id_contrato like '".$anho."'||'".$semestre."'
                    and tipo = 'J'
            )
                group by id_personal
                having decode(sign(sum(importe)),-1,0,sum(importe)) <> 0
            )
            group by aron.nivel_equiv(substr(aron.alumno_eap(id_personal),1,5),'001-
'||'".$anho."')
        )
        where UNI_ID like '530%'
        group by UNI_ID ";
```

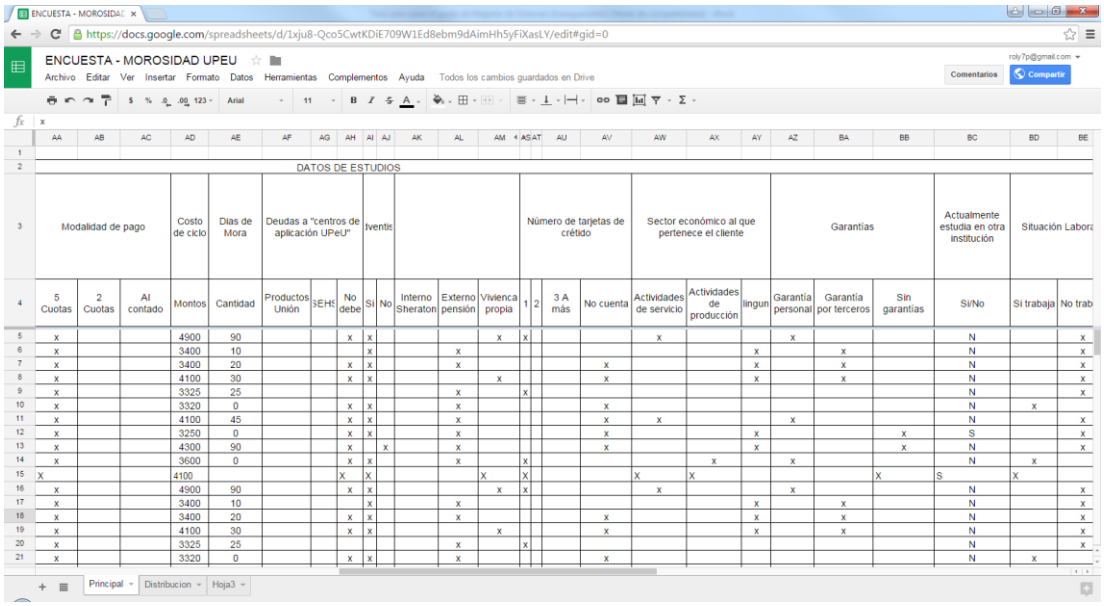
Anexo 6 - Encuesta realizado por vía web utilizando herramienta Google Docs.



ENCUESTA - MOROSIDAD UPEU

DATOS DEL ESTUDIANTE

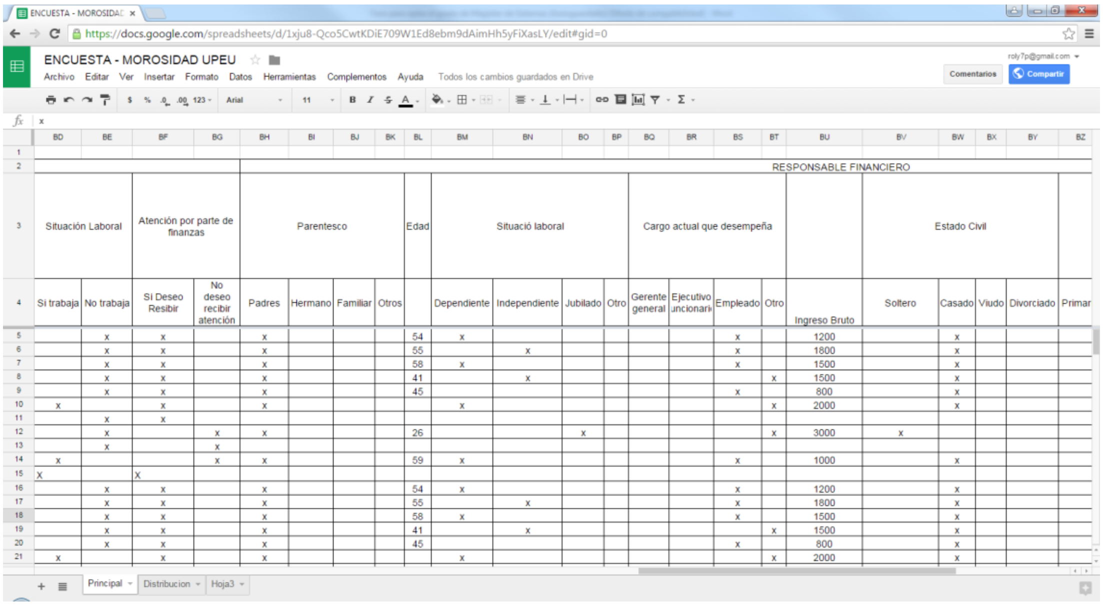
Edad	Sexo	Procedencia				Estado civil				Año de estudios					Financiamiento de estudios				Hermanos estudiando en la universidad				Ayuda por parte de la universidad			Modali
	F/M	País	Departamento	Provincia	Ciudad	Soltero	Casado	Viudo	Divorciado	1ero	2do	3ro	4to	5to	Ayuda de padres	Beca ayuda institucional	Autososten	Beca Ley	1	2	3 a más	No tiene	Beca de estudio	Beca de Feligresía	No tiene ayuda	5 Cuotas
26	M	Peru	Lima	Lima	Lima	x								x			x					x			x	x	
20	M	Peru	Lima	Lima	Lima	x						x			x						x			x		x	
22	F	Peru	Lima	Lima	Lima	x						x			x				x					x		x	
20	M	Peru	Lima	Lima	Lima	x						x			x							x		x		x	
25	M	peru	pasco	oxapampa	palcazu	x							x		x							x				x	
22	M	Perú	La libertad	Chepén	Chepén	x							x		x					x				x	x	x	
25	M	Peru	Lima	Lima	Lima	x								x			x					x		x		x	
26	M	Peru	Lima	Lima	Lima	x								x	x							x		x		x	
23	M	Perú	San Martín	Tocache	Tocache	x								x			x			x				x		x	
24	M	Peru	Lima	Lima	Lima	x								x	x							x		x		x	
25	m	peru	PUNO	PUNO	PUNO	X							X				X					X			X	X	
26	M	Peru	Lima	Lima	Lima	x								x			x					x			x	x	
20	M	Peru	Lima	Lima	Lima	x						x			x						x			x		x	
22	F	Peru	Lima	oyon	oyon	x							x		x				x					x		x	
20	M	Peru	Lima	Lima	Lima	x								x	x							x		x		x	
25	f	peru	cajamarca	cajamarca	cajamarca	x							x		x							x				x	
22	F	Perú	La libertad	La libertad	La libertad	x							x		x					x				x	x	x	



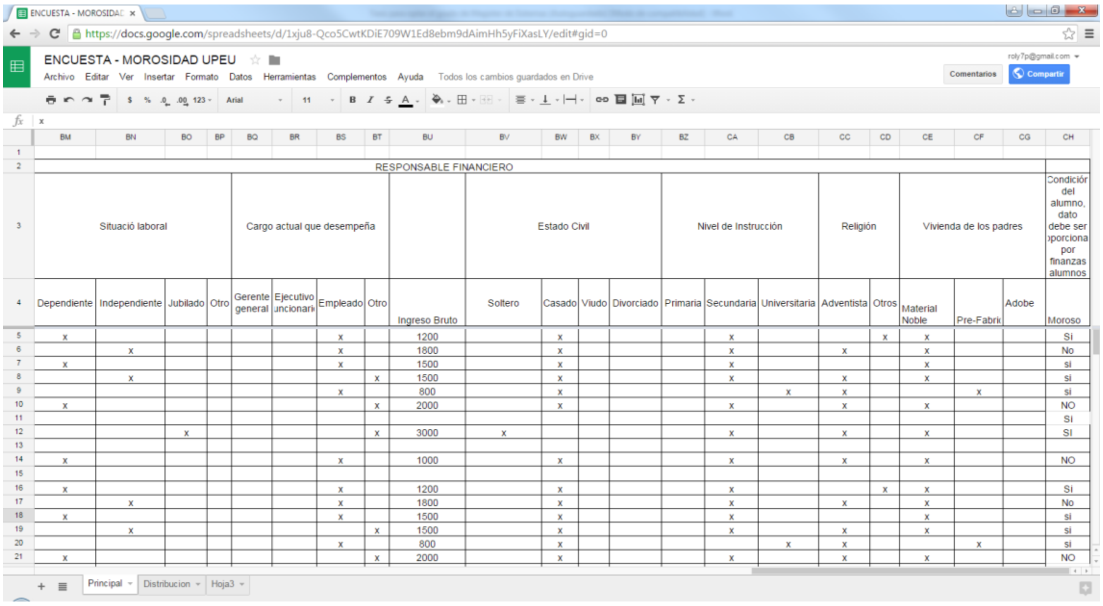
ENCUESTA - MOROSIDAD UPEU

DATOS DE ESTUDIOS

Modalidad de pago			Costo de ciclo	Dias de Mora	Deudas a "centros de aplicación UPeU"			Inventi					Número de tarjetas de crédito				Sector económico al que pertenece el cliente			Garantías			Actualmente estudia en otra institución	Situación Labora	
5 Cuotas	2 Cuotas	Al contado	Montos	Cantidad	Productos Unión	SEHS	No debe	SI	No	Interno Sheraton	Externo pensión	Vivienda propia	1	2	3 A más	No cuenta	Actividades de servicio	Actividades de producción	Ningun	Garantía personal	Garantía por terceros	Sin garantías	Si/No	Si trabaja	No trab
x			4900	90			x	x				x	x				x			x			N		x
x			3400	10				x			x								x		x		N		x
x			3400	20			x	x			x					x			x		x		N		x
x			4100	30			x	x				x				x			x		x		N		x
x			3325	25							x		x										N		x
x			3320	0			x	x			x					x							N	x	
x			4100	45			x	x			x					x	x			x			N		x
x			3250	0			x	x			x					x			x				S		x
x			4300	90			x		x		x					x			x			x	N		x
x			3600	0			x	x			x		x					x		x			N	x	
X			4100				X	X				X	X				X	X					X	S	X	
x			4900	90			x	x				x	x				x			x			N		x
x			3400	10				x			x								x		x		N		x
x			3400	20			x	x			x					x			x		x		N		x
x			4100	30			x	x				x				x			x		x		N		x
x			3325	25							x		x										N		x
x			3320	0			x	x			x					x							N	x	

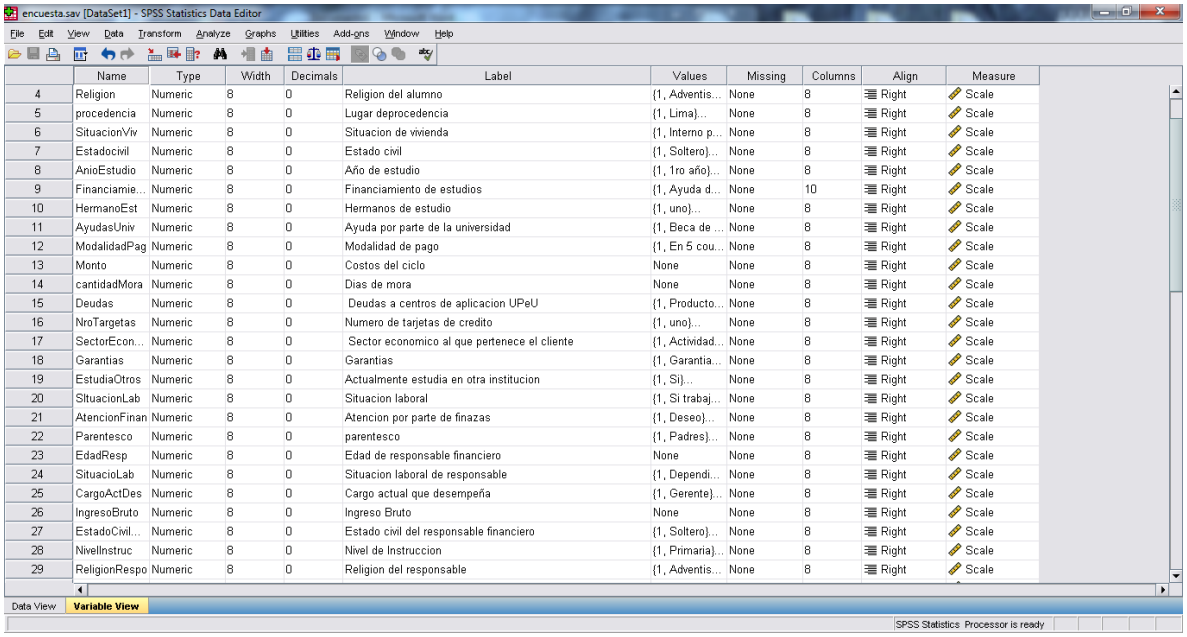


	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ
1																							
2					RESPONSABLE FINANCIERO																		
3	Situación Laboral		Atención por parte de finanzas		Parentesco				Edad	Situació laboral				Cargo actual que desempeña					Estado Civil				
4	Si trabaja	No trabaja	Si Deseo Resibir	No deseo recibir atención	Padres	Hermano	Familiar	Otros		Dependiente	Independiente	Jubilado	Otro	Gerente general	Ejecutivo uncionari	Empleado	Otro	Ingreso Bruto	Soltero	Casado	Viudo	Divorciado	Primar
5		x	x		x				54	x						x		1200		x			
6		x	x		x				55		x					x		1800		x			
7		x	x		x				58	x						x		1500		x			
8		x	x		x				41		x						x	1500		x			
9		x	x		x				45							x		800		x			
10	x		x		x					x							x	2000		x			
11		x	x																				
12		x		x	x				26			x					x	3000	x				
13		x		x																			
14	x			x	x				59	x						x		1000		x			
15	X		X																				
16		x	x		x				54	x						x		1200		x			
17		x	x		x				55		x					x		1800		x			
18		x	x		x				58	x						x		1500		x			
19		x	x		x				41		x						x	1500		x			
20		x	x		x				45							x		800		x			
21	x		x		x					x							x	2000		x			



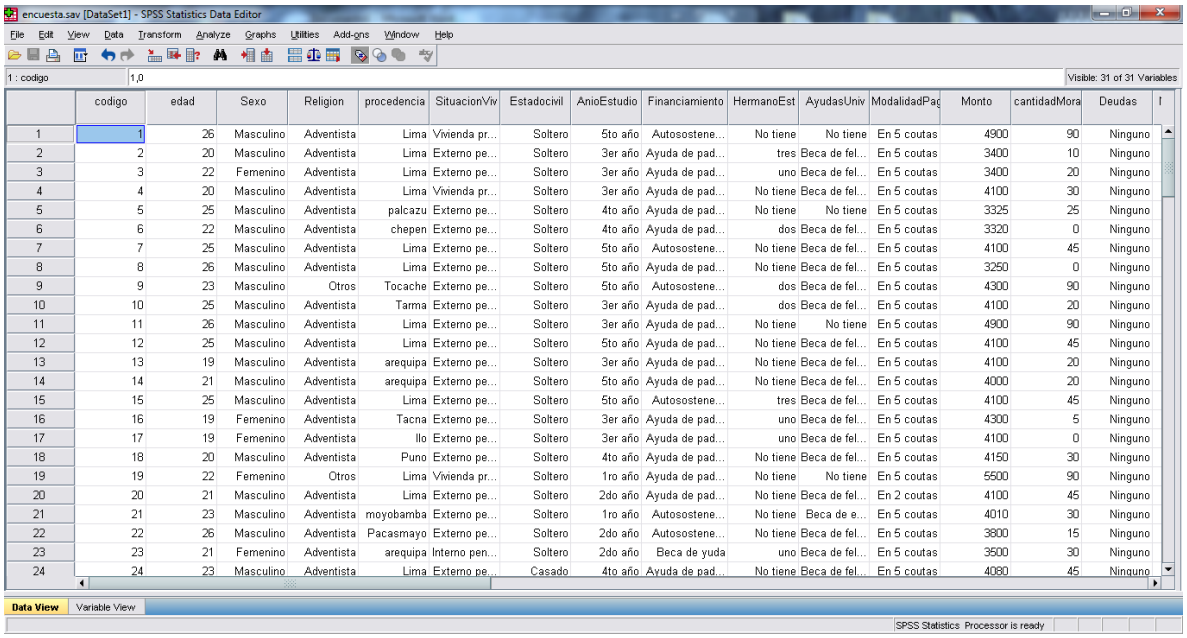
	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH
1																						
2	RESPONSABLE FINANCIERO																					
3	Situació laboral				Cargo actual que desempeña					Estado Civil				Nivel de Instrucción			Religión		Vivienda de los padres			Condición del alumno, dato debe ser proporciona por finanzas alumnos
4	Dependiente	Independiente	Jubilado	Otro	Gerente general	Ejecutivo uncionari	Empleado	Otro	Ingreso Bruto	Soltero	Casado	Viudo	Divorciado	Primaria	Secundaria	Universitaria	Adventista	Otros	Material Noble	Pre-Fabri	Adobe	Moroso
5	x						x		1200		x				x			x	x			Si
6		x					x		1800		x				x		x		x			No
7	x						x		1500		x				x				x			si
8		x						x	1500		x				x		x		x			si
9							x		800		x					x	x			x		si
10	x							x	2000		x				x		x		x			NO
11																						Si
12			x					x	3000	x					x		x		x			SI
13																						
14	x						x		1000		x				x		x		x			NO
15																						
16	x						x		1200		x				x			x	x			Si
17		x					x		1800		x				x		x		x			No
18	x						x		1500		x				x				x			si
19		x						x	1500		x				x		x		x			si
20							x		800		x					x	x			x		si
21	x							x	2000		x				x		x		x			NO

Anexo 7 - Llenado de datos a SPSS de encuesta realizado a los alumnos de la UPeU.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
4	Religion	Numeric	8	0	Religion del alumno	{1, Adventis...	None	8	Right	Scale
5	procedencia	Numeric	8	0	Lugar deprocedencia	{1, Lima}...	None	8	Right	Scale
6	SituacionViv	Numeric	8	0	Situacion de vivienda	{1, Interno p...	None	8	Right	Scale
7	Estadocivil	Numeric	8	0	Estado civil	{1, Soltero}...	None	8	Right	Scale
8	AnioEstudio	Numeric	8	0	Año de estudio	{1, 1ro año}...	None	8	Right	Scale
9	Financiamie...	Numeric	8	0	Financiamiento de estudios	{1, Ayuda d...	None	10	Right	Scale
10	HermanoEst	Numeric	8	0	Hermanos de estudio	{1, uno}...	None	8	Right	Scale
11	AyudasUniv	Numeric	8	0	Ayuda por parte de la universidad	{1, Beca de ...	None	8	Right	Scale
12	ModalidadPag	Numeric	8	0	Modalidad de pago	{1, En 5 cou...	None	8	Right	Scale
13	Monto	Numeric	8	0	Costos del ciclo	None	None	8	Right	Scale
14	cantidadMora	Numeric	8	0	Dias de mora	None	None	8	Right	Scale
15	Deudas	Numeric	8	0	Deudas a centros de aplicacion UPeU	{1, Producto...	None	8	Right	Scale
16	NroTargetas	Numeric	8	0	Numero de tarjetas de credito	{1, uno}...	None	8	Right	Scale
17	SectorEcon...	Numeric	8	0	Sector economico al que pertenece el cliente	{1, Actividad...	None	8	Right	Scale
18	Garantias	Numeric	8	0	Garantias	{1, Garantia...	None	8	Right	Scale
19	EstudiaOtros	Numeric	8	0	Actualmente estudia en otra institucion	{1, Si}...	None	8	Right	Scale
20	SituacionLab	Numeric	8	0	Situacion laboral	{1, Si trabaj...	None	8	Right	Scale
21	AtencionFinan	Numeric	8	0	Atencion por parte de finazas	{1, Deseo}...	None	8	Right	Scale
22	Parentesco	Numeric	8	0	parentesco	{1, Padres}...	None	8	Right	Scale
23	EdadResp	Numeric	8	0	Edad de responsable financiero	None	None	8	Right	Scale
24	SituacioLab	Numeric	8	0	Situacion laboral de responsable	{1, Dependi...	None	8	Right	Scale
25	CargoActDes	Numeric	8	0	Cargo actual que desempeña	{1, Gerente}...	None	8	Right	Scale
26	IngresoBruto	Numeric	8	0	Ingreso Bruto	None	None	8	Right	Scale
27	EstadoCivil...	Numeric	8	0	Estado civil del responsable financiero	{1, Soltero}...	None	8	Right	Scale
28	NivelInstruc	Numeric	8	0	Nivel de Instruccion	{1, Primaria}...	None	8	Right	Scale
29	ReligionRespo	Numeric	8	0	Religion del responsable	{1, Adventis...	None	8	Right	Scale

Aexo 8 - Preparacion de datos para la optension de resultados de morosidad de los alumnos.



encuesta.sav [DataSet1] - SPSS Statistics Data Editor

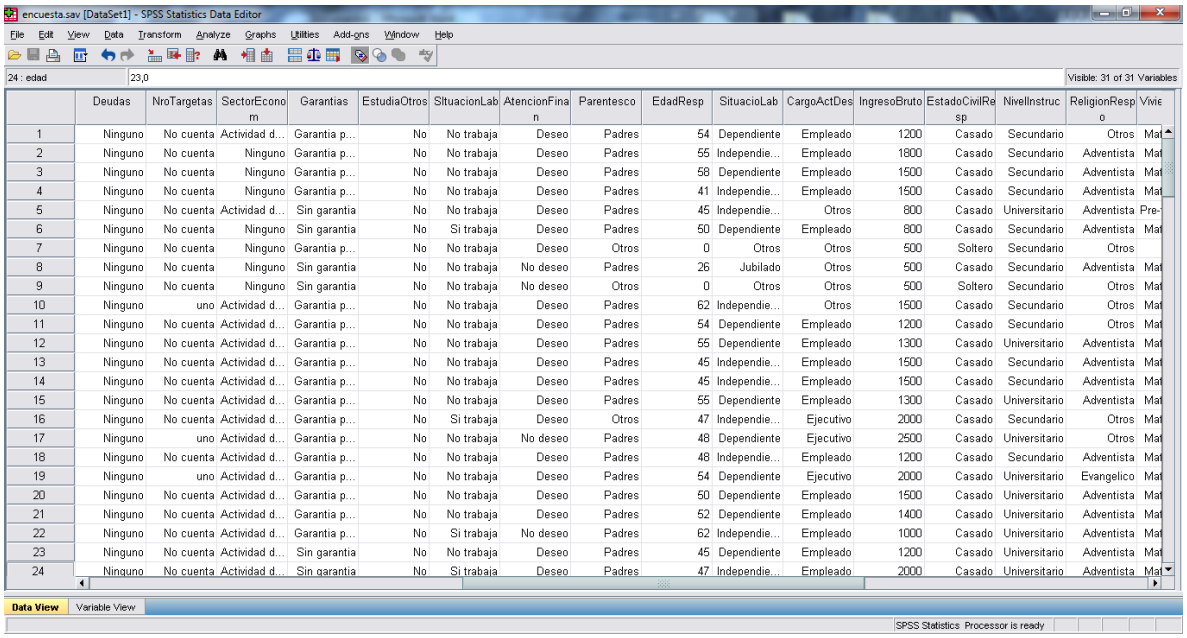
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : codigo 1,0 Visible: 31 of 31 Variables

	codigo	edad	Sexo	Religion	procedencia	SituacionViv	Estadocivil	AnioEstudio	Financiamiento	HermanoEst	AyudasUniv	ModalidadPag	Monto	cantidadMora	Deudas	I
1	1	26	Masculino	Adventista	Lima	Vivienda pr...	Soltero	5to año	Autosostene...	No tiene	No tiene	En 5 coutas	4900	90	Ninguno	
2	2	20	Masculino	Adventista	Lima	Externo pe...	Soltero	3er año	Ayuda de pad...	tres	Beca de fel...	En 5 coutas	3400	10	Ninguno	
3	3	22	Femenino	Adventista	Lima	Externo pe...	Soltero	3er año	Ayuda de pad...	uno	Beca de fel...	En 5 coutas	3400	20	Ninguno	
4	4	20	Masculino	Adventista	Lima	Vivienda pr...	Soltero	3er año	Ayuda de pad...	No tiene	Beca de fel...	En 5 coutas	4100	30	Ninguno	
5	5	25	Masculino	Adventista	palcazu	Externo pe...	Soltero	4to año	Ayuda de pad...	No tiene	No tiene	En 5 coutas	3325	25	Ninguno	
6	6	22	Masculino	Adventista	chepen	Externo pe...	Soltero	4to año	Ayuda de pad...	dos	Beca de fel...	En 5 coutas	3320	0	Ninguno	
7	7	25	Masculino	Adventista	Lima	Externo pe...	Soltero	5to año	Autosostene...	No tiene	Beca de fel...	En 5 coutas	4100	45	Ninguno	
8	8	26	Masculino	Adventista	Lima	Externo pe...	Soltero	5to año	Ayuda de pad...	No tiene	Beca de fel...	En 5 coutas	3250	0	Ninguno	
9	9	23	Masculino	Otros	Tocache	Externo pe...	Soltero	5to año	Autosostene...	dos	Beca de fel...	En 5 coutas	4300	90	Ninguno	
10	10	25	Masculino	Adventista	Tarma	Externo pe...	Soltero	3er año	Ayuda de pad...	dos	Beca de fel...	En 5 coutas	4100	20	Ninguno	
11	11	26	Masculino	Adventista	Lima	Externo pe...	Soltero	3er año	Ayuda de pad...	No tiene	No tiene	En 5 coutas	4900	90	Ninguno	
12	12	25	Masculino	Adventista	Lima	Externo pe...	Soltero	5to año	Ayuda de pad...	No tiene	Beca de fel...	En 5 coutas	4100	45	Ninguno	
13	13	19	Masculino	Adventista	arequipa	Externo pe...	Soltero	3er año	Ayuda de pad...	No tiene	Beca de fel...	En 5 coutas	4100	20	Ninguno	
14	14	21	Masculino	Adventista	arequipa	Externo pe...	Soltero	5to año	Ayuda de pad...	No tiene	Beca de fel...	En 5 coutas	4000	20	Ninguno	
15	15	25	Masculino	Adventista	Lima	Externo pe...	Soltero	5to año	Autosostene...	tres	Beca de fel...	En 5 coutas	4100	45	Ninguno	
16	16	19	Femenino	Adventista	Tacna	Externo pe...	Soltero	3er año	Ayuda de pad...	uno	Beca de fel...	En 5 coutas	4300	5	Ninguno	
17	17	19	Femenino	Adventista	Ilo	Externo pe...	Soltero	3er año	Ayuda de pad...	uno	Beca de fel...	En 5 coutas	4100	0	Ninguno	
18	18	20	Masculino	Adventista	Puno	Externo pe...	Soltero	4to año	Ayuda de pad...	No tiene	Beca de fel...	En 5 coutas	4150	30	Ninguno	
19	19	22	Femenino	Otros	Lima	Vivienda pr...	Soltero	1ro año	Ayuda de pad...	No tiene	No tiene	En 5 coutas	5500	90	Ninguno	
20	20	21	Masculino	Adventista	Lima	Externo pe...	Soltero	2do año	Ayuda de pad...	No tiene	Beca de fel...	En 2 coutas	4100	45	Ninguno	
21	21	23	Masculino	Adventista	moyobamba	Externo pe...	Soltero	1ro año	Autosostene...	No tiene	Beca de e...	En 5 coutas	4010	30	Ninguno	
22	22	26	Masculino	Adventista	Pacasmayo	Externo pe...	Soltero	2do año	Autosostene...	No tiene	Beca de fel...	En 5 coutas	3800	15	Ninguno	
23	23	21	Femenino	Adventista	arequipa	Interno pen...	Soltero	2do año	Beca de yuda	uno	Beca de fel...	En 5 coutas	3500	30	Ninguno	
24	24	23	Masculino	Adventista	Lima	Externo pe...	Casado	4to año	Ayuda de pad...	No tiene	Beca de fel...	En 5 coutas	4080	45	Ninguno	

Data View Variable View

SPSS Statistics Processor is ready



encuesta.sav [DataSet1] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

24 : edad 23,0 Visible: 31 of 31 Variables

	Deudas	NroTargetas	SectorEconom	Garantias	EstudiaOtros	SituacionLab	AtencionFinan	Parentesco	EdadResp	SituacioLab	CargoActDes	IngresoBruto	EstadoCivilResp	NivelInstruc	ReligionResp	Vivie
1	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	54	Dependiente	Empleado	1200	Casado	Secundario	Otros	Mat
2	Ninguno	No cuenta	Ninguno	Garantia p...	No	No trabaja	Deseo	Padres	55	Independie...	Empleado	1800	Casado	Secundario	Adventista	Mat
3	Ninguno	No cuenta	Ninguno	Garantia p...	No	No trabaja	Deseo	Padres	58	Dependiente	Empleado	1500	Casado	Secundario	Adventista	Mat
4	Ninguno	No cuenta	Ninguno	Garantia p...	No	No trabaja	Deseo	Padres	41	Independie...	Empleado	1500	Casado	Secundario	Adventista	Mat
5	Ninguno	No cuenta	Actividad d...	Sin garantia	No	No trabaja	Deseo	Padres	45	Independie...	Otros	800	Casado	Universitario	Adventista	Pre-
6	Ninguno	No cuenta	Ninguno	Sin garantia	No	Si trabaja	Deseo	Padres	50	Dependiente	Empleado	800	Casado	Secundario	Adventista	Mat
7	Ninguno	No cuenta	Ninguno	Garantia p...	No	No trabaja	Deseo	Otros	0	Otros	Otros	500	Soltero	Secundario	Otros	
8	Ninguno	No cuenta	Ninguno	Sin garantia	No	No trabaja	No deseo	Padres	26	Jubilado	Otros	500	Casado	Secundario	Adventista	Mat
9	Ninguno	No cuenta	Ninguno	Sin garantia	No	No trabaja	No deseo	Otros	0	Otros	Otros	500	Soltero	Secundario	Otros	Mat
10	Ninguno	uno	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	62	Independie...	Otros	1500	Casado	Secundario	Otros	Mat
11	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	54	Dependiente	Empleado	1200	Casado	Secundario	Otros	Mat
12	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	55	Dependiente	Empleado	1300	Casado	Universitario	Adventista	Mat
13	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	45	Independie...	Empleado	1500	Casado	Secundario	Adventista	Mat
14	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	45	Independie...	Empleado	1500	Casado	Secundario	Adventista	Mat
15	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	55	Dependiente	Empleado	1300	Casado	Universitario	Adventista	Mat
16	Ninguno	No cuenta	Actividad d...	Garantia p...	No	Si trabaja	Deseo	Otros	47	Independie...	Ejecutivo	2000	Casado	Secundario	Otros	Mat
17	Ninguno	uno	Actividad d...	Garantia p...	No	No trabaja	No deseo	Padres	48	Dependiente	Ejecutivo	2500	Casado	Universitario	Otros	Mat
18	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	48	Independie...	Empleado	1200	Casado	Secundario	Adventista	Mat
19	Ninguno	uno	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	54	Dependiente	Ejecutivo	2000	Casado	Universitario	Evangelico	Mat
20	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	50	Dependiente	Empleado	1500	Casado	Universitario	Adventista	Mat
21	Ninguno	No cuenta	Actividad d...	Garantia p...	No	No trabaja	Deseo	Padres	52	Dependiente	Empleado	1400	Casado	Universitario	Adventista	Mat
22	Ninguno	No cuenta	Actividad d...	Garantia p...	No	Si trabaja	No deseo	Padres	62	Independie...	Empleado	1000	Casado	Universitario	Adventista	Mat
23	Ninguno	No cuenta	Actividad d...	Sin garantia	No	No trabaja	Deseo	Padres	45	Dependiente	Empleado	1200	Casado	Universitario	Adventista	Mat
24	Ninguno	No cuenta	Actividad d...	Sin garantia	No	Si trabaja	Deseo	Padres	47	Independie...	Empleado	2000	Casado	Universitario	Adventista	Mat

Data View Variable View

SPSS Statistics Processor is ready

Anexo 9 - Comprobación de la puntuación media del coeficiente de la morosidad del grupo alumnos que no paga las cuotas según la fecha establecida en su contrato (después de los 25 días).