# FACULTAD DE INGENIERÍA

# CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS

## TESIS

"APLICACIÓN DE MINERÍA DE DATOS PARA PRONOSTICAR EL RIESGO DE MOROSIDAD DE LOS ESTUDIANTES DE LA UNIVERSIDAD AUTÓNOMA DEL PERÚ"

## PARA OBTENER EL TÍTULO DE

INGENIERO DE SISTEMAS

## AUTOR(ES)

ALENSTER YONEL CORDOVA VALDIVIA

KAREN IVETT TORRES JURADO

## ASESOR

MG. JOSE LUIS HERRERA SALAZAR

**LIMA, PERÚ, JULIO DE 2018**

## DEDICATORIA

A mis padres que siempre me brindaron su apoyo y creyeron en mí y ser el pilar fundamental en todo lo que soy, académicamente y como de la vida.

Alenster Yonel Córdova Valdivia

A Dios por haberme guiado dándome las fuerzas para continuar y culminar esta etapa de mi vida, gracias a su infinita bondad y amor.

A mis padres que siempre me brindaron su apoyo y creyeron en mí y ser el pilar fundamental en todo lo que soy, académicamente y como de la vida.

A mi familia en general, porque me han brindado su apoyo incondicional y por estar conmigo en los buenos y malos momentos.

Karen Ivett Torres Jurado

## AGRADECIMIENTOS

La presente tesis, ha sido realizado gracias a Dios, por habernos acompañado y guiado hasta este momento tan importante de nuestra formación profesional, por ser nuestra fortaleza en los momentos de debilidad y brindarnos una vida llena de aprendizajes y experiencias.

A nuestros padres por apoyarnos en todo momento, por los valores que nos han inculcado, por su comprensión, amor y ayuda en los momentos difíciles de nuestras vidas.

A la Universidad Autónoma del Perú y también nos gustaría agradecer al cuerpo de docentes de la Facultad de Ingeniería porque todos aportaron con un granito de arena a nuestra formación profesional.

De igual manera agradecer al Ing. Jose Luis Herrera Salazar, al Ing. Pretell, al Ing. Camacho cuyos conocimientos impartidos nos permitieron contar con las herramientas necesarias para la culminación de esta tesis.

Son muchas las personas que han formado parte de nuestra vida profesional a las que nos encantaría agradecer su amistad, consejos, apoyo, ánimo y compañía en los momentos más difíciles de nuestras vidas. Darles gracias por formar parte de nosotros, por todo lo que nos han brindado y por todas sus bendiciones.

# RESUMEN

En épocas de crisis, controlar y gestionar la morosidad pasa a ser una de las principales preocupaciones de las empresas.

El presente proyecto plantea la aplicación de modelo predictivo para clasificar calidad de pago de los estudiantes en la Universidad Autónoma del Perú

La finalidad de implementar una aplicación en la Universidad Autónoma del Perú es contar con una herramienta tecnológica que detecte a un alumno de ser un posible moroso, logrando la prevención de la situación de dificultad financiera en la empresa, con un enfoque en cómo evitar la morosidad y a su vez realizar un correcto control y seguimiento de los impagos.

**Palabras clave:** Modelo Predictivo, Minería de Datos, Metodología CRISP – DM, Técnica del árbol, Patrones, Controlar y gestionar morosidad, Base de Datos.

# ABSTRACT

In times of crisis, controlling and managing delinquency becomes one of the main concerns of companies.

This project proposes the application of a predictive model to classify students' payment quality at the Universidad Autónoma del Perú

The purpose of implementing an application in the Universidad Autónoma del Perú is to have a technological tool that detects a student of being a possible defaulter, achieving the prevention of the situation of financial difficulty in the company, with a focus on how to avoid delinquency and in turn, make a correct control and follow-up of defaults.



**Keywords:** Predictive Model, Data Mining, CRISP - DM Methodology, Tree Technique, Patterns, Control and manage delinquency, Database.

# ÍNDICE DE CONTENIDO

## ÍNDICE DE TABLAS

# ÍNDICE DE FIGURAS

# INTRODUCCIÓN

El presente trabajo de investigación tuvo como objetivo principal utilizar la técnica del árbol siguiendo el esquema de la metodología Crisp-DM, para clasificar calidad de pago de los estudiantes en la universidad Autónoma del Perú

En la actualidad existen muchas excusas por la cual el cliente no cancela su deuda, ya que, si no se cobra dicha deuda, no habría liquidez y la empresa debe saber distinguir entre el cliente que tiene un problema real y el cliente que simula un problema.

Ante este hecho, el gestor debe evaluar una serie de características para decidir dentro de los parámetros que se pueda manejar, si le otorga un plan de facilidades, un tiempo de espera, ya que para ello se tiene que evaluar diferentes situaciones como por ejemplo antigüedad, comportamiento histórico de pago, si es cliente nuevo, cual es la deuda total, es correcta la decisión de reprogramar la deuda.

En esta evaluación es muy importante contar con la historia completa del cliente. Como así también, contar con una aplicación predictiva, que nos permita diferenciar si es un cliente potencialmente moroso o no. Este tipo de sistema nos facilitará la tarea de tomar las decisiones y especialmente las excepciones.

El presente proyecto consiste en aplicar un modelo predictivo enfocado a mejorar el pronóstico ante los posibles morosos en la Universidad Autónoma del Perú, Esto conlleva a que las personas que toman decisiones en la organización tengan una herramienta capaz de agilizar este proceso, de forma que las decisiones que se tomen sean las más acertadas y que esto se refleje en rentabilidad para la Universidad.

Las limitaciones encontradas en la fase de desarrollo de la aplicación fueron las limitadas respuestas de los usuarios, generando así retrasos en la recolección de información, encuestas y/o entrevistas.

Con el propósito a que se haga más entendible la presente tesis, ha sido dividido en 5 Capítulos, cuyo contenido son los siguientes:

En el **capítulo I:** Planteamiento Metodológico. - Se detalla todo referente al planteamiento metodológico, pues involucra la definición del problema, justificación, nivel de investigación, objetivos, hipótesis, variables e indicadores, diseño de investigación y los métodos de recolección de datos.

El Marco Referencial definido en el **capítulo II:** Se detalla antecedentes, teniendo como referencia tesis, libros y artículos; y la parte teórica de la tesis, la validación del marco teórico relacionando con las metodologías y modelos que se están usando para el desarrollo de la tesis.

Se tiene el **capítulo III:** Desarrollando el Sistema Web. - Esta es la parte más importante de la tesis ya que se describe la parte de desarrollo del Sistema Web usando las metodología y etapas ya definida en el marco teórico

En el **capítulo IV:** Análisis e Interpretación de los Resultados. - Se realiza la prueba empírica para la recopilación, análisis e interpretación de los resultados obtenidos. En el primer lugar se describe la población y muestra, seguidamente el tipo de muestra, nivel de confianza. También se muestra el análisis de los datos preprueba y post prueba. Los datos se muestran en tablas las cuales al término de este capítulo son analizados y seguidamente se realiza la contratación de la hipótesis.

Y para terminar el **capítulo V:** Conclusiones y Recomendaciones**.** - Se muestran las conclusiones y recomendaciones de la tesis

Al final se presenta la referencia bibliográfica, apéndices y el glosario de términos

# CAPÍTULO I

# PLANTEAMIENTO DEL PROBLEMA

## 1.1 EL PROBLEMA

### 1.1.1 Descripción de la Realidad Problemática

**PERÚ**

Actualmente en el Perú, según estudios de la INEI, se cuenta con un total de 142 universidades registradas de las cuales el 57% están institucionalizadas, este porcentaje equivale a 76 centros de estudio, por el contrario, el otro porcentaje que equivale a 64 universidades tienen un permiso temporal de funcionamiento.

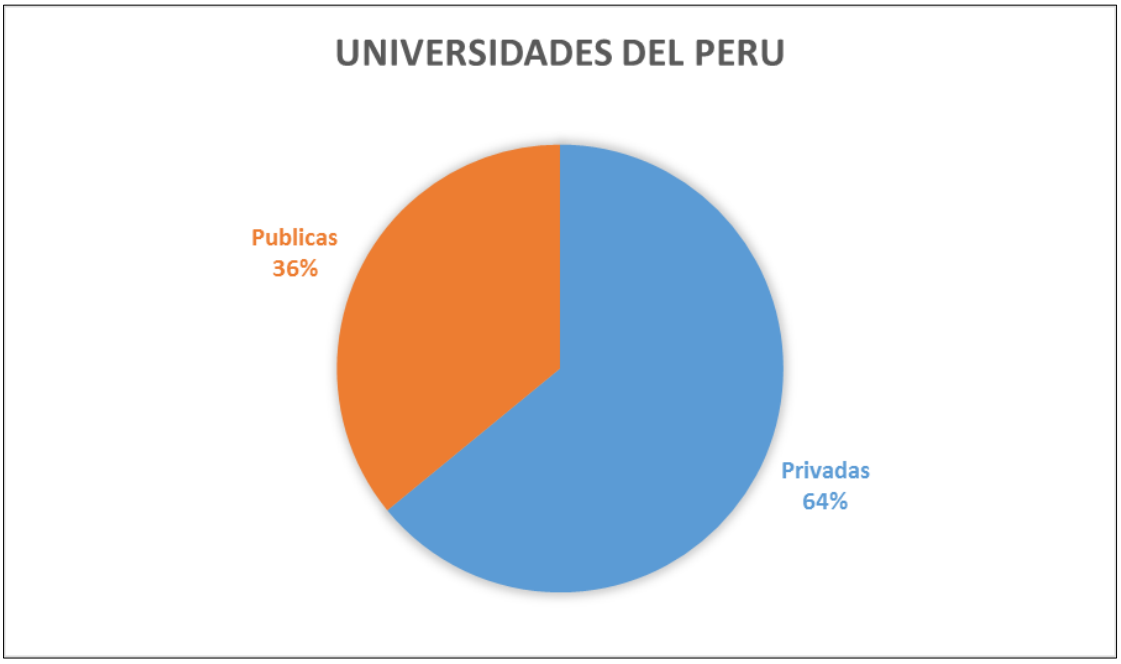


*Figura 1.* Superintendencia Nacional de Educación Superior. Adaptado de "Superintendencia Nacional de Educación Superior" por (Inei, 2012)

Se denota constantemente sus deudas, ya sea los pagos que tienen que realizar o cualquier sector que tengan que ver con transacciones monetarias, ya que no es muy difícil de ver a personas con deudas en bancos, casas de compra y venta, así entre otros establecimientos por causa de no haber realizado en su debido tiempo, un estudio de la persona que se le ha de brindar el bien o servicio, son muy pocos los ejemplos de establecimientos que hacen un seguimiento a una persona que ha de pedir un crédito o algo referente a ello para que más adelante no tenga los problemas que se ven día con día en nuestra sociedad.

## SECTOR EDUCATIVO

Los centros de pagos de las universidades en este caso las privadas, se ven inmersas en problemas ya descritos anteriormente con respecto al dinero y esto conlleva a la desconfianza que generan porque si bien en claro hay personas de bajos recursos que piden prorrogas para poder seguir estudiando, hay otros que simplemente no desean pagar por diferentes motivos. Estos casos generan la desconfianza a prestar ayuda a personas que si la necesitan.

Con este proyecto de tesis se plantea de forma tecnológica aplicar la técnica del árbol para el modelo predictivo que nos permita hacer un seguimiento de los nuevos alumnos y de los que ya están dentro de la universidad y poder establecer los potenciales casos de personas que no cumplan con los reglamentos monetarios establecidos por la universidad (Universidad Autónoma del Perú, 2016).

## CENTRO DE PAGOS

Es un centro en el cual se realiza la labor de la cobranza de la mano de la asistenta social según se vea el caso, en esta área se ve los distintos pagos desde la matrícula de nuevo ingresante hasta los cursos extracurriculares que se llevan a cabo en la universidad.

Ubicándonos en la realidad, ponemos de referencia al centro de finanzas de la universidad Autónoma del Perú que mediante un proceso cualitativo empírico, obtiene los diversos motivos que se ven inmersos en el no cumplimiento de las cuotas, por diversos hechos que lo vuelven un alumno moroso, siguiendo un protocolo de 3 fases, desde el mismo alumno el centro de finanzas y el área de cuentas corrientes.

La investigación se centra en el área de finanzas ve la manera de evitar que el alumno muy aparte no sea moroso, tenga la posibilidad de refinanciar su deuda, este proceso se hace mediante un dialogo, utilizando el conocimiento de la experiencia de la o las personas que estén en ese momento para poder dejar claro los puntos que se van a tratar, de esta manera enfatizar que tienen que pagar, y a su vez evitar la deserción del alumnado por falta de pagos sea por el motivo que sea (Universidad Autónoma del Perú, 2016).

### 1.1.2 Descripción del problema

En épocas de crisis, controlar y gestionar la morosidad pasa a ser una de las principales preocupaciones de las empresas.

Una de las problemáticas que tiene la Universidad Autónoma del Perú es el poco control de los alumnos morosos, esto se da por la falta de algún método que determine cuál es el porcentaje que el alumno sea un posible moroso, ya que la finalidad de implementar un modelo predictivo en la Universidad Autónoma del Perú es contar con una herramienta tecnológica que detecte a un alumno de ser un posible moroso, logrando la prevención de la situación de dificultad financiera en la empresa, con un enfoque en cómo evitar la morosidad y a su vez realizar un correcto control y seguimiento de los impagos, ubicándonos en el área de finanzas nos damos un claro ejemplo en la realidad que se vive respecto a este tema, ya que la tasa de morosos es sumamente alta y esto conlleva a la factibilidad de desarrollo del proyecto.

**UBICACIÓN**

La Universidad Autónoma del Perú se encuentra ubicada en la panamericana Sur, Km 16.3, Villa El Salvador, Lima – Perú


*Figura 2.* Ubicación de la Universidad Autónoma del Perú. Adaptado de "Ubicación de la Universidad Autónoma del Perú" por (Google Maps, 2016)

La Universidad Autónoma del Perú es una institución dedicada a formar personas y profesionales íntegros, responsables y competitivos activamente en la sociedad y de la ciencia, contribuyendo a una sociedad a través de una educación de calidad basada en propuestas innovadoras en el marco de principios y valores universales.

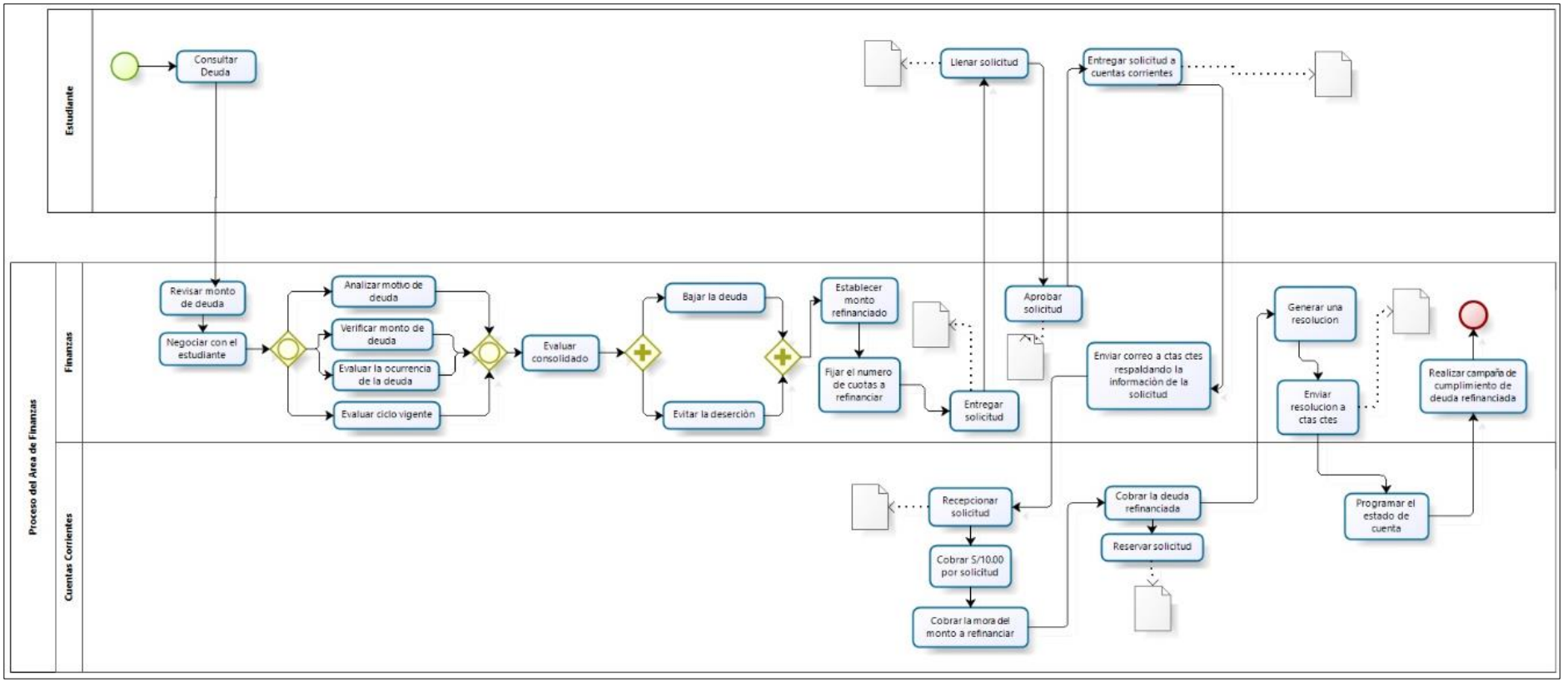


*Figura 3*. Flujograma del Proceso.

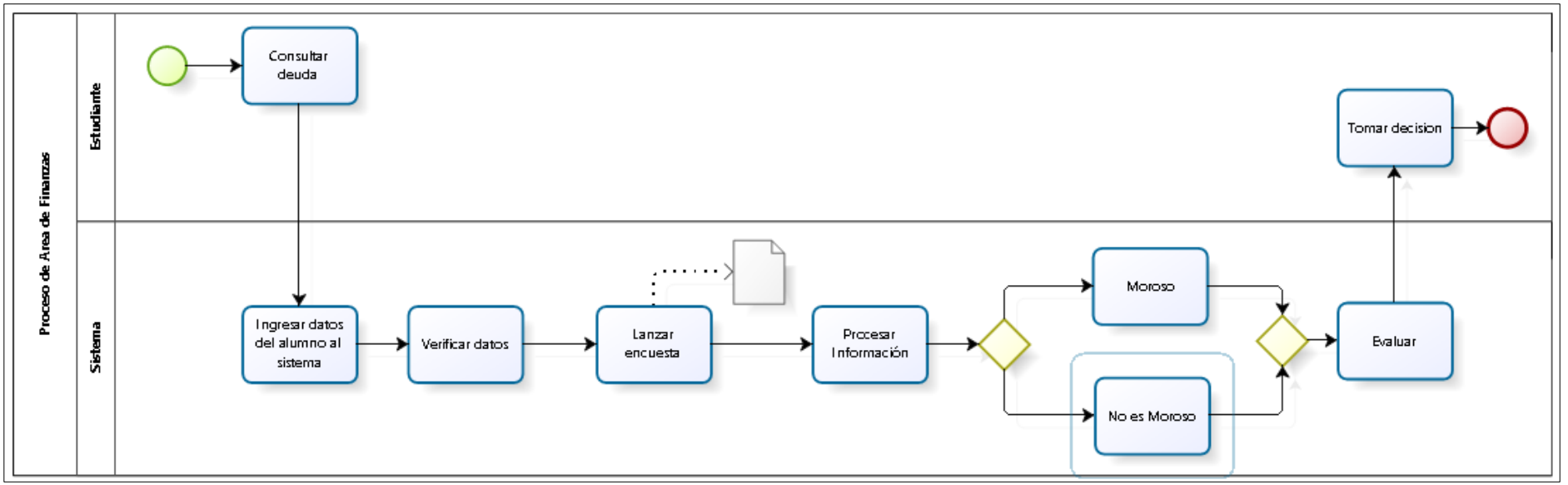


*Figura 4*. Flujograma del Proceso del Sistema

Observando el panorama del diagrama anterior, a continuación, se describen cinco problemas los cuales fueron motivos para el desarrollo de este proyecto de tesis.

Entre los principales problemas tenemos:

- Tiempo para verificar datos del alumno
- Tiempo para registrar datos
- Tiempo para verificar los alumnos morosos
- Tiempo para procesar información
- Tiempo para verificar porcentaje de alumnos morosos
- Tiempo para generar exactitud de reportes de los alumnos morosos
- Nivel de Satisfacción

Tabla 1
*Cuadro Comparativo entre la Situación Actual (AS – IS) y la Situación (TO – BE)*

| Indicador | Datos Pre – Prueba (Promedio) |
|---|---|
| **Tiempo para determinar perfiles de los alumnos morosos** | 60 minutos/alumno |
| **Tiempo para predecir al alumno moroso** | 15 minutos/solicitud |
| **Tiempo para verificar porcentaje de alumnos morosos** | 30 minutos/ |
| **Tiempo para generar exactitud de reportes de los alumnos morosos** | 20 minutos/ |
| **Nivel de Satisfacción** | Regular |

Tabla 2
*Cuadro Comparativo entre la Situación Actual (AS – IS) y la Situación (TO – BE)*

| AS – IS | TO - BE |
|---|---|
| **Tiempo inadecuado de registrar al alumno** | Tiempo apropiado para registrar al alumno |
| **Demasiado tiempo en registrar datos del moroso** | Tiempo apropiado para registrar datos del moroso |

| | |
|---|---|
| **Demasiado tiempo para procesar información** | Tiempo apropiado para procesar información |
| **Demasiado tiempo para verificar porcentajes de alumnos morosos** | Tiempo adecuado para verificar porcentajes de alumnos morosos |
| **Nivel de Satisfacción baja** | Regular |

### 1.1.3 Enunciado del Problema

¿En qué medida un aplicativo de minería de datos basado en la técnica de árboles de decisión facilitara la predicción de la morosidad de los estudiantes de la Universidad Autónoma del Perú?

## 1.2 TIPO Y NIVEL DE INVESTIGACION

### 1.2.1 Tipo de Investigación

**Aplicada**

El presente trabajo aplicará la técnica del árbol para predecir a los posibles morosos en potencia, apoyándonos de la metodología Crisp-Dm y de las prácticas del Business Intelligence, enfocadas en el Data Minig.

### 1.2.2 Nivel de Investigación

**Explicativa**

Se orienta a establecer las causas que originan un fenómeno determinado. Se trata de un tipo de investigación cuantitativa que descubre el por qué y el para qué de un fenómeno.

**Experimental**

En la actualidad no existe un modelo de predicción en el área de finanzas, entonces a través de esta investigación de tipo experimental, aplicamos nuestro modelo predictivo para evaluar de mejor manera la toma de decisiones.

**Correlacionar**

Mediremos el grado de relación en un ambiente controlado, por lo tanto, se puede decir que en este caso aplicaremos el modelo predictivo para mejorar el pronóstico de elección de posibles morosos en la Universidad Autónoma del Perú.

## 1.3 JUSTIFICACION DE LA INVESTIGACIÓN

El desarrollo de esta investigación se basa en el estudio y puesta en funcionamiento de un modelo predictivo que permitirá al personal llevar de mejor manera el control de los posibles morosos y así mejorar la selección que es reportado por los colaboradores del área de Finanzas en la Universidad Autónoma del Perú

### 1.3.1 Teoría

La minería de datos es un campo de la estadística y las ciencias de la computación referido a la secuencia de procesos que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.

El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, supone aspectos de gestión de datos y de bases de datos, de procesamiento de datos, del modelo y de las consideraciones de inferencia, de métricas de Intereses, de consideraciones de la Teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea.

### 1.3.2 Practica

En este estudio se tiene mapeado el poder predecir quienes serán morosos en la universidad teniendo en cuenta que esa laborar se tiene conocimiento empírico por lo cual no se tiene un diagnostico exacto de quienes serían los potenciales morosos por lo cual esta técnica ayudara a poder determinar quién será el deudor a futuro.

Se optimizará la labor del área de finanzas ya que este sistema ayudará a poder anticipar a los posibles morosos según sus comportamientos.

### 1.3.3 Metodología

Si bien es cierto la metodología que se ha usado para este trabajo es la metodología Crisp-Dm se tienen 2 metodologías más que acompañando a la elegida son las predominantes.

Dichas metodologías son:

KDD: Es una metodología propuesta por Fayyad en 1996, propone 5 fases: Selección, preprocesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo.

SEMMA: Es el acrónimo a las cinco fases: (Sample, Explore, Modify, Model, Assess) La metodología es propuesta por SAS Institute Inc, la define como: "proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos"

Aun así, se quedó con la metodología CRISP-DM ya que es una de las más completas según se requiere el estudio.

## 1.4 OBJETIVO DE LA INVESTIGACION

### 1.4.1 Objetivo General

Determinar en qué medida el uso de una aplicación de minería de datos influenciara en el pronóstico de riesgo de morosidad de los estudiantes de la Universidad Autónoma del Peru-2016.

### 1.4.2 Objetivos Específicos

1. Determinar en qué medida un aplicativo de minería de datos basado en la técnica de árboles reducirá el tiempo para verificar el alumno moroso de la Universidad Autónoma del Perú en el año 2016.

2. Determinar en qué tiempo un modelo predictivo clasificaría a los posibles alumnos morosos de la Universidad Autónoma del Perú en el año 2016.

3. Determinar en qué medida un aplicativo de minería de datos basado en la técnica de árboles reducirá el tiempo para verificar el alumno moroso de la Universidad Autónoma del Perú en el año 2016.

4. Determinar en qué tiempo se realizaría el estudio cualitativo del comportamiento de las personas de la Universidad Autónoma del Perú en el año 2016.

5. Determinar una investigación del comportamiento de las personas para generar conocimiento sobre los riesgos en los otorgamientos de facilidades de la Universidad Autónoma del Perú en el año 2016.

6. Determinar los beneficios que se obtendrán al aplicar el modelo predictivo en la Universidad Autónoma del Perú en el año 2016.

7. Determinar los beneficios de usar Data Mining para la obtención del modelo predictivo en la Universidad Autónoma del Perú en el año 2016.

## 1.5 HIPOTESIS

### Hipótesis General

El uso de un modelo predictivo basado en la técnica de árboles influye en la predicción de riesgo al otorgar facilidades económicas a los estudiantes de la Universidad Autónoma del Perú

### Hipótesis Específicos

1. La aplicación de un modelo predictivo respaldara los veredictos de finanzas respecto al tema de los morosos en la Universidad Autónoma del Perú.

2. La aplicación de un modelo predictivo otorgara patrones para poder identificar a los estudiantes morosos en la Universidad Autónoma del Perú.

## 1.6 VARIABLES E INDICADORES

### 1.6.1 Variable Independiente

#### A) Indicadores

**Variable Independiente: Aplicación de Minera de Datos.**

Tabla 3
*Este indicador permite conocer su existencia o ausencia*

| **Indicador: Presencia – Ausencia** |
|---|
| Descripción: Cuando es NO, es porque no se ha desarrollado un Modelo predictivo usando Data Mining para poder clasificar para la Universidad autónoma del Perú y aún se encuentra en la situación del problema. Cuando es SI, es que se ha desarrollado un Modelo predictivo usando Data Mining para la Universidad autónoma del Perú, esperando buenos resultados**.** |

### 1.6.2 Variable Dependiente

**Variable Independiente:** Predicción de la morosidad

Tabla 4
*Proceso de Finanzas de posibles morosos de la Universidad Autónoma del Perú*

| Indicador | Descripción |
|---|---|
| **Tiempo para verificar datos del alumno** | Es el tiempo que se utiliza para verificar historia del alumno |
| **Tiempo para verificar alumno moroso** | Es el tiempo que se realiza para verificar si el alumno es un posible moroso |
| **Índice de Riesgo** | Es la factibilidad de que haya riesgo de morosidad |
| **Tiempo para Determinar el riesgo de morosidad de un estudiante** | Es el tiempo que se realizara para procesar la información del estudiante |
| **Tiempo para predecir que alumnos incurrirán en morosidad** | Es el tiempo que se tomara para poder dar la predicción del posible moroso |

## B) Operacionalización

### Variable Independiente

Tabla 5
*Modelo predictivo usando Data Mining para poder clasificar*

| Indicador | Índice |
|---|---|
| **Presencia – Ausencia** | No, Si |

### Variable Dependiente

### Predicción de la morosidad

Tabla 6
*Proceso de Finanzas de posibles morosos de la Universidad Autónoma del Perú.*

| Indicador | Índice | Unidad de Medida | Unidad observación |
|---|---|---|---|
| **Tiempo para verificar datos del alumno** | [5...10 | Minutos/verificar alumno | Registro Manual |

| | | | |
|---|---|---|---|
| **Tiempo para verificar alumno moroso** | [10…15 | Minutos/registrar alumno | Registro Manual |
| **Índice de Riesgo** | [15…25 | Minutos/alumno moroso | Registro Manual |
| **Tiempo para Determinar el riesgo de morosidad de un estudiante** | [10…15 | Minutos/procesar información | Registro Manual |
| **Tiempo para predecir que alumnos incurrirán en morosidad** | [10…15 | Minutos/procesar información | Registro Manual |

## 1.7 LIMITACIONES DE LA INVESTIGACIÓN

**Acceso:**

Es muy limitada la información del alumnado para poder realizar los estudios pertinentes

**Tiempo:**

Es muy limitado para poder entrevistar a la variedad de realidades que podemos encontrar

## 1.8 DISEÑO DE LA INVESTIGACIÓN

**Gc O1 X O2**

**Dónde:**

- Gc = Grupo de Control: Es un grupo de control al que no se aplicará el estímulo (Modelo predictivo).
- X = Modelo predictivo: Estímulo o condición experimental.
- O1 = Datos de la Pre-Prueba.

- O2 = Datos de la Post-Prueba para los indicadores de la Variable Dependiente una vez implementado el Módulo predictivo: Mediciones Post-Prueba del grupo de control.
- = Falta de estímulo o condición experimental.

**Descripción:**

Se trata de escoger de forma intencional de un grupo experimental (Ge), al que se aplicará un modelo predictivo (X), el cual se les aplica a trabajadores del Área de finanzas de la Universidad Autónoma del Perú (Gc); y se realiza pruebas mediante el módulo predictivo (O2), conformado de manera intencional por trabajadores del Área de Finanzas de la en la Universidad Autónoma del Perú, donde no se le aplicará el estímulo, sirviendo sólo como grupo de control; en forma simultánea ya que se le aplica una prueba.

Los dos grupos están constituidos de forma intencional pero representativa estadísticamente. Tanto en ausencia como en presencia del modelo predictivo.

## 1.9 TÉCNICAS E INSTRUMENTO PARA LA RECOLECCIÓN DE INFORMACIÓN

### 1.9.1 Técnicas de la Investigación de Campo.

Tabla 7
*Técnicas e Instrumentos de la Investigación de Campo*

| TÉCNICAS | INSTRUMENTOS |
|---|---|
| **1. Observaciones Directa**<br>• **Estructurada**<br>• **No Participante** | • Diario de Campo<br>• Fichas de Observación |
| **2. Aplicación de Encuestas** | • Encuestas (documento) |

### 1.9.2 Instrumentos de la Investigación de Campo.

Tabla 8
*Técnica e Instrumentos de la Investigación de Campo*

| TÉCNICAS | INSTRUMENTOS |
|---|---|
| **Revisión de:** | |
| • **Libros**<br>• **Revistas**<br>• **Artículos**<br>• **Manuales**<br>• **Tesis**<br>• **Información Web**<br>• **Periódicos** | • Laptops<br>• Tablet<br>• Memorias USB<br>• Impresiones<br>• Blocks<br>• Fotocopias<br>• Smartphone |

# CAPÍTULO II
# MARCO TEÓRICO

## 2.1 ANTECEDENTES DE LA INVESTIGACIÓN

**A) Autor:**

- Mauricio Moreno Echevarría
- Víctor Ovalle Retamal

**Título:** Aplicación de un modelo predictivo de fuga de clientes utilizando data mining en VTR Globalcom S.A. zona sur.

**Tipo de tesis:** Pregrado

**Año:** 2011

**Correlación:**

Esta investigación tiene como objetivo aplicar un modelo de predicción de fuga de clientes en la compañía VTR Global COM S.A. Zona Sur.

Se basa en la metodología y herramienta de Data Mining para determinar el modelo de predicción de fuga de clientes. Se utilizó la base de datos de la compañía, de la cual se extrajo información demográfica de clientes y también de características de servicios contratados.

El tiempo elegido para desarrollar esta investigación se comprende desde enero de 2009 hasta junio de 2011. Ya que, en conjunto con analistas de la compañía, se eligieron las variables que podrían ser más influyentes. Luego mediante el proceso KDD (Knowledge Discovery in Data bases) se logró limpiar y transformar las variables que luego fueron incluidas en las dos técnicas específicas utilizadas de Data Mining.

La primera técnica utilizada fue Análisis Cluster (o Análisis de Conglomerados), permite generar perfiles de clientes fugados, en las cuales se establecen características tales como: edad, antigüedad de servicios, niveles de deuda, y otras dieciséis variables relevantes para el estudio.

La segunda técnica utilizada fue Regresión Logística Multivariante, permite estudiar la relación entre la variable dependiente y las variables independientes, aquí se realiza el análisis univariantes, bivariantes y evaluación de posibles interacciones o modificaciones de tipo efecto y/o confusión, luego se realizó la validación con meses

de prueba (abril 2011 – junio 2011) en la cual se pudo verificar el modelo con respecto a los meses que se utilizó para construir de estos mismos (enero 2009 – marzo 2011). En general ambos modelos tienen un nivel de acierto global superior a un 70%.

En conclusión, se generó un plan de acción en base a Marketing Relacional, ya que las debilidades presentes con respecto a la generación en post de retención y captación de clientes. Esta estrategia propone capacitar al personal de ventas y atención al cliente logrando generar una herramienta de apoyo que permita fidelizar al corto plazo de mejor manera a los clientes de ambas plazas. (Moreno y Ovalle, 2011)

**Punto de vista**

Esta tesis nos ayuda a ver otra opción de hacer los modelos predictivos, ya que usa la técnica de clústeres o agrupamientos, obteniendo de estar forma grupos de comportamientos y/o patrones para poder hacer su comparativa, de esta forma poder llegar a obtener los resultados deseados, en la tesis se ve los tipos de análisis que han usado para poder definir los parámetros a estudiar, estas tesis nos ayudan a tener una visión más amplia de los múltiples tipos de análisis predictivos que hay.

**B) Autor:** Erwin Sergio Fischer Angulo

**Título**: Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios.

**Tipo de Tesis:**

**Año:** 2012

**Correlación:**

Resumen: La deserción universitaria se ha convertido en un gran problema a ser investigado. La tasa de deserción llego a ser uno de los principales indicadores. Ya que el invertir más tiempo en diagnóstico de las causas de la deserción con metodologías que permitan predecir con mayor efectividad.

El objetivo consiste en investigar una metodología, la cual permita identificar de una forma automática los estudiantes con mayor riesgo de deserción.

Para realizar la implementación se trabajó con la metodología CRISP-DM y con modelos de Redes Neuronales, Árbol de Decisión y Cluster K-medianas y poder analizar el comportamiento de los estudiantes.

La exactitud de los modelos es calculada a partir el conjunto de datos de prueba, los cuales indican que ningún modelo predictivo arrojó resultados positivos, mediante esto se analizó que es muy probable que los datos no eran suficientemente confiables. Dado que dentro de los límites de este trabajo era imposible conseguir datos fidedignos y completos, esta tesis propone una metodología para enfrentar estudios de minería de datos educativa donde se eviten los problemas descritos. (Fischer, 2012)

**Punto de vista**

Esta tesis nos da un enfoque un poco más amplio ya que se muestran diferentes técnicas para hacer la predicción de los alumnos que van a desertar, usan la metodología crisp-dm, la cual nos ayudó a revisar los puntos en nuestro desarrollo muy aparte de esto nos dio un mejor enfoque de la técnica que vamos a usar, a su vez nos mostró diversas formas de darle solución ya que se ve el uso de la técnica de las redes neuronales, Clústeres kmedians para poder desarrollar el modelo, gracias a estos puntos de vista pudimos recopilar la información necesaria para ajustar los parámetros de nuestro modelo predictivo.

**C) Autor: Jesús Walter Salinas Flores**

**Título:** Reconocimiento de patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación cart.

**Tipo de tesis**:

**Año**: 2005

**Correlación:**

Resumen: En los sistemas financieros es importante despejar el riesgo crediticio cuando se le concede un crédito a un cliente para un producto determinado. El presente trabajo de investigación tiene como objetivo encontrar una secuencia de comportamientos de la morosidad a partir de la información recopilada al momento de solicitar un producto financiero y a su vez dar a conocer una metodología practica

para este campo que es la técnica de los Árboles de Regresión CART la cual se aplica en situaciones donde se tienen un grupo de datos de individuos en los que se han medido variables productoras o independientes y una variable de clasificación o de criterio que define el grupo al que cada individuo pertenece; y se quiere encontrar un conjunto de reglas de decisión que permitan explicar la clasificación existente y utilizar estas reglas para poder clasificar a un nuevo individuo, esta técnica tiene la ventaja de una alta potencia en sus estimadores.

Se presenta el marco teórico empezando con una caracterización del sistema financiero nacional. Luego se dan algunas referencias sobre el riesgo de crédito y se dan las bases teóricas sobre el reconocimiento de patrones. Finalmente se presenta la metodología del árbol de clasificación CART que es objetivo del presente trabajo y una definición conceptual de términos.

La metodología de la investigación está referenciada a la descripción del producto crediticio y la estrategia para la prueba de hipótesis. Para aplicar el algoritmo de árbol de clasificación CART se ha trabajado con un producto crediticio del sistema financiero nacional que otorga préstamos en dólares a personas naturales. (Salinas, 2005)

**Punto de vista**

En esta tesis podemos observar de mejor manera la técnica que vamos usar ya que se usa la técnica del árbol, con la diferencia que en este proyecto se está usando la clasificación CART esto nos da un énfasis distinto ya que se está usando regresión, al usar regresión ya deja de ser un modelo supervisado, lo que no deja como enseñanza es la forma de usar la técnica de árboles dándonos un mejor ejemplo de cómo plantear este diseño para poder hacer las predicciones de manera más fluida y concisa.

**Autor (s): Dr. PAULINO E. MALLO; Universidad Nacional de Mar del Plata**

**Título del Paper: ANÁLISIS DE LA MOROSIDAD TRIBUTARIA DE LAS EMPRESAS APLICANDO TÉCNICAS BORROSAS Y ESTADÍSTICAS. EL CASO DE MAR DEL PLATA (Mallo, 2010)**

**Revista: Anales de las Jornadas Internacionales de Estadística**

**Volumen (Edición): Volumen I**

**pág. – pág.: 1-9**

---

**Estado del arte que hace el autor**

El propósito del presente trabajo es identificar el comportamiento de los indicadores contables de distintas empresas de pequeña y mediana envergadura del medio socio productivo de la ciudad de Mar del Plata que puedan incidir en la clasificación de la morosidad en el pago de tributos, sean nacionales, provinciales o municipales, a través de la aplicación de lógica difusa. Para ello se trabajó con una muestra de empresas representativas de distintos sectores productivos y comerciales de la ciudad de Mar del Plata, a su vez considerando empresas de distintos tamaños y participaciones en el mercado, construyendo ratios contables indicativos de rentabilidad, solvencia, liquidez, rotación y endeudamiento de las mismas, a partir de la información recopilada en los estados contables de las respectivas firmas correspondientes al último ejercicio fiscal cerrado. Las empresas seleccionadas son PyMEs marplatenses de entre 10 y 100 empleados dedicadas al comercio minorista, industrias de sectores clave para la economía local - como la alimenticia, la construcción y vinculadas con el puerto - y servicios estratégicos en la ciudad, como la educación, la gastronomía, el hotelería y el turismo. El aporte de la lógica difusa al estudio de dichos indicadores es de dar una mejor precisión de las reglas de comportamiento para el análisis de la morosidad en el pago de tributos por parte de las empresas marplatenses y su relación con la situación económico-financiera de las mismas, representada a través de diversas ratios contables. La definición de reglas de comportamiento mediante proposiciones lingüísticas favorecerá una comprensión conceptual de la realidad económica financiera de la empresa inserta en los procesos decisorios.

## 1. Descripción del aporte del autor

El objetivo de la investigación es analizar la incidencia de las distintas ratios indicativas de la realidad económica y financiera de las empresas, tales como endeudamiento, rentabilidad, solvencia, liquidez y rotación, en la morosidad en el pago de tributos de jurisdicción nacional, provincial y municipal. Para ello se aplicó modelos difusos, donde las variables explicativas escogidas se traducen en una predicción de los valores de la variable dependiente a través de una función matemática y de un conjunto de reglas difusas. Los modelos basados en la lógica difusa permiten obtener un conjunto de reglas, que en nuestro caso de estudio evidencian el comportamiento tributario de las firmas establecidas como unidades de análisis. El propósito del modelo es doble, por un lado, caracterizar a través de las reglas el comportamiento de las unidades de análisis, y por otro obtener un valor clarificativo para el endeudamiento fiscal, permitiendo a los expertos humanos un mejor análisis de las decisiones de pago de los tributos por parte de las entidades analizadas.

## 2. Proceso para resolver el problema considerado por el autor

Con el propósito de dar más relevancia a la información económica y financiera sobre las PyMEs marplatenses y su comportamiento en el pago de tributos, se accedió a los últimos estados contables de 80 empresas de pequeña y mediana envergadura que desarrollan su actividad en la ciudad de Mar del Plata (de los cuales el 20% se utilizaron para testear el modelo). En este trabajo se presentan los resultados preliminares de una investigación sobre el perfil moroso de las PyMES de nuestra ciudad. Con el objetivo de que la muestra sea relativamente homogénea respecto del tamaño de las empresas, se descartaron aquellas que poseían menos de 10 empleados y más de 150. También se consideró que estuvieran representados los principales sectores comerciales y productivos de la ciudad. Así, se obtuvieron los datos de empresas comerciales, de la industria alimenticia, de la construcción, dedicadas a los servicios turísticos, hoteleros y gastronómicos, empresas agrícola-ganaderas, metalúrgicas, y de servicios de salud, educación y transportes. Sobre los estados contables presentados por dichas empresas.

**3. Métricas que el autor usa y resultado que obtiene. Comentar (los resultados son mejores respecto a otros)**

La metodología para la construcción del sistema implica: a. definir las variables de entrada - independientes - y salida - dependiente. A partir de la selección de las variables de entrada y salida se definen las categorías lingüísticas, los conjuntos difusos y las funciones de pertenencia asociadas.

A cada indicador considerado en el análisis se le asocian niveles lingüísticos a las variaciones de medida que experimenta, estas categorías pueden ser: leve, moderado, medio, alto y superior. Cada uno de estos términos lingüísticos define un conjunto difuso en sí mismo que se representa a través de una función de pertenencia μ – valor numérico en que se expresa la variable lingüística –. La función de pertenencia elegida para representar cada categoría lingüística se corresponde con un grado de membresía entre 0 y 1. La función de pertenencia utilizada para los diferentes conjuntos difusos de los indicadores elegidos es gaussiana. La construcción del modelo se basó en una clusterización difusa y en una implicación Sugeno. La clusterización difusa se basa en la identificación de centros de clúster, de acuerdo con la densidad de los puntos definidos como centros y agrupando el resto de los datos según sus distancias a dichos centros, en una función de minimización. Así, cada clúster define un conjunto difuso para cada variable.

**4. Conclusiones**

La provisión de reglas difusas que muestran el comportamiento de las PyMES marplatenses, provee a los expertos de una herramienta de apoyo para la toma de decisiones que logra objetividad y uniformidad en la formulación de criterios para la evaluación del cumplimiento de pago de tributos. Esta información resultaría beneficiosa para los análisis de morosidad de los clientes efectuados por las entidades recaudadoras fiscales tanto en el ámbito nacional como en el provincial y municipal. También sería de utilidad para las entidades financieras que otorguen créditos o financien a futuro diferentes actividades de estas empresas en la determinación de la posibilidad del pago de las obligaciones. Es de destacar que el modelo preliminar de investigación sobre el desarrollo tributario de las PyMES marplatenses descripto en este trabajo, puede ser complementado por la experiencia de una serie de expertos durante el proceso de clasificación de contribuyentes por parte del fisco. (Mallo, 2010)

**Punto de Vista:**

El articulo nos da el énfasis necesario para poder conocer sobre el método predictivo de lógica difusa y métodos borrosos que viene a tener correlación con los temas de inteligencia artificial, dando énfasis en la predicción de morosidad en el sector financiero, utilizando muestras de distintas empresas dedicadas a distintos rubros, la ayuda que nos da el artículo es la metodología que utiliza para poder precisar las reglas de entrenamiento que utilizaran en el algoritmo que será el motor del modelo predictivo a su vez del entrenamiento que pueda tener el algoritmo para mejorar la precisión y clasificación de los distintos casos que se ven presente durante el estudio.

**Author (s): Vargas, Hovanna; Ccapa, Lesly**

**Títle of paper: Modelo de Árboles de decisión para pronosticar la morosidad de los alumnos de la Universidad Peruana Unión. (Vargas y Ccapa, 2011) Jornal: Revista de Busisness Intelligence**

**Volume (issue): Volumen I**

**pág. – pág.: 26 - 32**

---

## 1. Estado del arte que hace el autor

Para el pronóstico de morosidad de los alumnos de la Universidad Peruana Unión se construyó y validó una encuesta para la recolección de datos, la cual fue tomada a los alumnos de las diferentes escuelas de la Universidad.

Árboles de Decisión. (Según Departamento de Informática Universidad Nacional de San Luis (UNSL) San Luis. Argentina octubre de 2006.) El aprendizaje de árboles de decisión es un método que ha sido utilizado en numerosas tareas de aprendizaje inductivo. Es un método de aproximación de funciones robusto a la presencia de datos erróneos y es capaz de aprender expresiones disyuntivas.

Existe toda una gama de algoritmos de aprendizaje de árboles de decisión que incluye a algoritmos muy conocidos como ID3, ASSISTANT y C4.5. Esta familia de algoritmos, referenciada a veces como TDIDT (Top-Down Induction of Decision Trees) se identifican por buscar un margen en la hipótesis completamente expresivo

que evita las dificultades de los espacios de hipótesis restringidos. Su sesgo inductivo es un sesgo de preferencia por árboles pequeños sobre árboles grandes.

Crossland menciona que los árboles de decisión son herramientas excelentes para ayudar a realizar elecciones adecuadas entre muchas posibilidades. Su estructura permite seleccionar una y otra vez diferentes opciones, que pueden tener diferentes alternativas que al ser exploradas pueden ser una posible decisión.

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión proporcionada por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos.

El algoritmo genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con una columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

## 2. Descripción del aporte del autor

Actualmente en el mundo cada año aumentan las entidades crediticias, el mercado es cada vez más competitivo; por lo tanto, una entidad crediticia debe ejercer control efectivo sobre el proceso de evaluación de sus clientes con el fin de otorgarle o negarle el crédito solicitado.

La cartera de crédito al consumo implica el manejo de un gran número de clientes. Las entidades financieras requieren procesar un gran número de solicitudes de crédito, por tanto, es importante que la administración deba conocer el comportamiento de sus clientes.

El riesgo de crédito es el tipo de riesgo más importante al que debe hacer frente cualquier entidad financiera. Un indicador del riesgo crediticio es el nivel de morosidad de la entidad, es decir, la proporción de su cartera que se encuentra en calidad de incumplimiento.

La Universidad se encuentra con un porcentaje critico de alumnos morosos en las diferentes facultades y escuelas, lo cual está conllevando a una gran preocupación en

el directorio general, debido a que está alterando el pago a tiempo a su personal, los recursos destinados al mantenimiento de su infraestructura y las ganancias a la institución, esto sucede debido que no se está controlando la morosidad en los alumnos en el departamento de finanzas.

Es por esto la necesidad de desarrollar un modelo que muestre el pronóstico de los alumnos morosos de la Universidad Peruana Unión y apoye a la toma de decisiones al área de Finanzas; lo cual es una ventaja en el negocio.

Con esta investigación se contribuirá a concretar la automatización de las actividades relacionadas con la morosidad de la universidad, ayudará a discernir el comportamiento crediticio de los alumnos de la universidad. Los datos detallados de los alumnos morosos permitirán tomar un mejor control del problema de la morosidad y aplicar las medidas para evitarlas en el futuro.

## 3. Proceso para resolver el problema considerado por el autor

**Fases del modelo CRISP-DM:**

Comprensión del Negocio: establecimiento de los objetivos del negocio, evaluación de la situación, generación del plan del proyecto.

Comprensión de los Datos: recopilación inicial de datos; descripción, exploración y verificación de la calidad de datos.

Preparación de Datos: selección de datos construcción e integración de los datos.

Modelado: Aplicación de las técnicas de minería de datos; selección y diseño de la evaluación, construcción del modelo de árboles de decisión y la evaluación respectiva.

Evaluación: Evaluación de los resultados del modelo, de acuerdo a las necesidades del negocio y establecimiento de los pasos a seguir.

Despliegue: Integración el resultado del modelo a las actividades del negocio, planificación, monitorización y revisión del proyecto.

## ANÁLISIS DEL MODELO UTILIZADO

Análisis de la eficiencia del modelo de árboles de decisión.

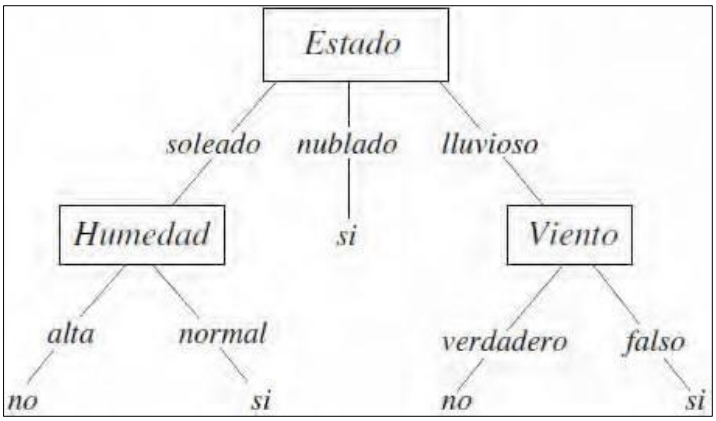


*Figura 5.* Árbol de decisión. Adaptado de "Árbol de decisión" por (Vargas y Ccapa, 2011)

Este ejemplo de árboles de decisión se trata de decidir si vamos a jugar tenis dependiendo, de las condiciones atmosféricas siguientes: nubosidad, humedad y viento. (cielo=soleado, temperatura=caliente, humedad=alta, viento=fuerte)

Tabla 9
*Clasificación de tipo de clima*

| Día | Cielo | Temperatura | Humedad | Viento | Jugar tenis |
|---|---|---|---|---|---|
| **D1** | Soleado | Alta | Alta | Débil | No |
| **D2** | Soleado | Alta | Alta | Fuerte | No |
| **D3** | Lluvioso | Alta | Alta | Débil | Sí |
| **D4** | Lluvioso | Media | Alta | Débil | Sí |
| **D5** | Lluvioso | Fría | Normal | Débil | Sí |
| **D6** | Lluvioso | Fría | Normal | Fuerte | No |
| **D7** | Lluvioso | Fría | Normal | Fuerte | Sí |
| **D8** | Soleado | Media | Alta | Débil | No |
| **D9** | Soleado | Fría | Normal | Débil | Sí |
| **D10** | Lluvioso | Media | Normal | Débil | Sí |
| **D11** | Soleado | Media | Normal | Fuerte | Sí |
| **D12** | Lluvioso | Media | Alta | Fuerte | Sí |

| | | | | | |
|---|---|---|---|---|---|
| **D13** | Lluvioso | Alta | Normal | Débil | Sí |
| **D14** | Lluvioso | Media | Alta | Fuerte | No |

Entropía ([9+,5-])= - (9/14) log2 (9/14) -(5/14) log2 (5/14) = 0’940.

Supongamos que S es un conjunto de entrenamiento con 14 ejemplos

A.9 ejemplos positivos y 5 negativos ([9+,5-]). B. Unos de los atributos, Viento, puede tomar los valores Débil y Fuerte.

Tabla 10

*La ganancia de información que obtenemos si clasificamos los 14 ejemplos según el atributo Viento*

| | **Positivos** | **Negativos** |
|---|---|---|
| **Débil** | 6 | 2 |
| **Fuerte** | 3 | 3 |

$$\text{Entropía (S)} - \sum_{ve} Valores(A) \frac{|S_v|}{|S|} Entropia(S_v)$$

$$= \text{Entropía (S)} - \frac{|8|}{14} Entropia(S\_Débil) - \frac{|6|}{14} Entropia(S\_Fuerte)$$

$$= 0.811 - \frac{|8|}{14} 1.00$$

*Figura 6.* Valores y Entropía

=0.940

=0.048

Ganancias (S, Humedad) Ganancia (S, Viento)

=0.940 – (7/14) * 0.985 =0.940 – (8/14)*0.811

-(7/14) * 0.592 -(6/14) * 1.00

=0.151 =0.048

## 4. Conclusiones

Usando el algoritmo de árboles de decisión se detectaron patrones para los morosos y los no morosos. El principal patrón detectado es que el ingreso económico de los padres sea menor a 1600 nuevos soles, que no tenga ayuda de la universidad, que la

situación de los padres es independiente (jubilado), y que tengan un monto creditico 1201 y 1600 nuevos soles con la universidad. Este patrón caracteriza al 33,75% de los alumnos son morosos.

Aplicando, este tipo de investigaciones en este el rubro, que es el brindar créditos educativos, estaremos previniendo futuros endeudamientos y falta de pagos en las instituciones. Además, sabremos a qué tipo de clientes podremos otorgarle un crédito de acuerdo a las variables definidas de acuerdo al análisis realizado.

El estudio realizado posee un margen de error, no se puede afirmar del todo que un cliente que no posee la cantidad requerida de ingresos tienda a ser deudor, pero posee los indicios. Además, los que poseen una elevada cantidad de ingreso económico, tiendan a ser morosos por más que el modelo de estudio diga lo contrario.

**Punto de vista:**

El paper nos brinda una mejor perspectiva del tema a tratar de la tesis ya que abarcar en su mayoría casi todos los puntos que se deben ver en el tema a tratar, dándonos un claro ejemplo de cómo orientarnos, brindándonos una mejor perspectiva, al momento de usar las técnicas de la predicción y la forma de como entrenar al algoritmo que hará el trabajo de darnos las respuestas para poder valorar en qué nivel de morosidad esta la persona que pase la evaluación pertinente.

**Autor (s): Chuquival Samuel, Galindos Jaime, Maquera Saúl**

**Títle of paper: Modelo de redes neuronales para mejorar el pronóstico del comportamiento del alumno en el cumplimiento del pago de sus armadas, concernientes a un crédito aprobado por el área de finanzas Alumnos de la Universidad Peruana Unión (Chuquival, Galindo y Maquera, 2012)**

**Journal: Revista de Busisness Intelligence**

**Volume (issue): Volumen I pág. – pág.: 12 – 17**

---

## 1. Estado del arte que hace el autor

El riesgo en dar un crédito es la principal fuente de problemas en los entes financieros. El motivo es la cartera de créditos como es el activo más importante y con mayor participación en una cooperativa que desarrolla actividades financieras.

Los centros bancarios tienen un gran paradigma el cual consiste en el ofrecer crédito a bajo costo, y de manera oportuna, frente a la competencia, el éxito del negocio

depende del normal retomo de las operaciones crediticias financiadas, con un bajo nivel de riesgo, e incrementando el número de alumnos con el beneficio del crédito para que continúen sus estudios.

Como función principal, el Área de Finanzas de la Universidad Peruana Unión, gestiona los créditos y cobranza.

Tomando como referencia a un 50% de la población general, encontrando el 90.91% de No morosos, ajustándose al modelo ideal del 90.91% con una probabilidad de predicción del 99.03%

**Riesgo Crediticio**

Se define al riesgo crediticio como el miedo de alguien en devolver una cantidad de dinero que adeuda.

Es el riesgo de que un cliente o contraparte no pueda o no quiera cumplir con un compromiso que ha pactado con un miembro o miembros de una Institución.

En conciso, el riesgo crediticio es el que se concede a los clientes y en el concurren a su vez: riesgo de solvencia del cliente, riesgo jurídico y riesgo técnico instrumental. La naturaleza de cada uno de estos riesgos, su identificación, su minimización y su control, constituyen el objetivo principal del departamento de créditos.

## 2. Descripción del aporte del autor

**MODELO DE RNA BACK - PROPAGATION**

**Backpropagation**

La propagación hacia atrás de errores o retro-propagación (del inglés backpropagation) es un algoritmo de aprendizaje supervisado que se usa para entrenar redes neuronales artificiales.

El algoritmo de retro propagación analiza un margen de objetos (imágenes, colores, letras del alfabeto, datos u otros) de características similares uno por uno y va formando un patrón que representa a todos los objetos analizados, el error de aprendizaje de la RNA se minimizara siempre y cuando el algoritmo a analizarse y la cantidad sea adecuada a los objetos.

## 3. Proceso para resolver el problema considerado por el autor

**Metodología CRISP-DM**

El Programa ESPRIT desarrollo un proyecto en el año 2006. “Industria de la Cruz Norma Proceso de minería de datos” mencionando la metodología CRISP-DM como un conjunto de tareas realizadas 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.

El método CRISP-DM consta de 7 FASES (pasos dentro del proceso): comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y desarrollo.

Las fases del proyecto de Minería de acuerdo a lo establecido por la metodología CRISP-DM se conectan entre ellas de forma iterativa durante el desarrollo del proyecto.

Para la fase de "Comprensión del negocio" se analizó la problemática del área de finanzas alumnos y se establecieron las estrategias de solución, objetivos, requerimientos, restricciones, riesgos con sus respectivas contingencias, el cronograma y los costos.

En el período de "Comprensión de los datos" se impulso un sondeo para la colección de los datos los cuales fueron explorados y validados en el Software Estadístico SPSS. En la etapa de "Preparación de los datos" se realizó un blanqueo de los datos que no lograron ser validados en el período preliminar, con el software para explotación de datos SQL Server 2008 de @Microsoft.

En el período de "Modelado" se seleccionó el modelo de Malla Neuronal Back–Propagation posteriormente de formar un cotejo con otros modelos, conjuntamente se seleccionó el instrumento de estudio para explotación de datos SQL Server 2008.

En la fase "Evaluación" se obtuvieron los resultados arrojados por la herramienta ya mencionada y por último la fase de "Desarrollo" es la implementación de la diligencia de RNA entrenada testeada y puesta en elaboración.

**4. Métricas que el autor usa y resultado que obtiene. Comentar (los resultados son mejores respecto a otros)**

El modelo Back Propagation cimentado en redes Neuronales de Business Intelligence del SQL Server 2008 para esta indagación fue determinado por las variables representadas. Los datos de entrada fueron:AyudaporU, Estu-diaenotra, Financiamiento, IngresoPadres, Moda-lidaPago, MontoCredito, Montos, Peligro, Secto-reconomico, Sitlaboral, SitlaboralPadre y Tarje-tasCred.

El campo de entrada fue Carácter y la de predicción Riesgo. Del proceso para la edificación del modelo BACKPROPAGATION basado en redes neuronales podemos afirmar que:

Se realizó el desarrollo de amaestramiento de las variables de acceso con relación a la versátil de pronóstico con un porcentaje de casos al 90%.
La clasificación del proceder del alumno en el cumplimiento del desembolso de sus armadas viene a ser dada por 2 categorías: No moroso y Moroso. El modelo Back - Propagation generado obtuvo una calificación de 0.97 y representando una profecía con una posibilidad de confiabilidad del 93.29%. Sobre el estudio con el algoritmo Back - Propagation cimentado en redes neuronales: Se analizaron 750 casos de la muestra general obtenida. Sobre estos alumnos, los padres con ingresos menores 1200 Nuevos Soles poseen superior valor de peligro crediticio. Finalmente, los alumnos con padres no jubilados aún representan infracción en el desembolso de créditos.

## 5. Conclusiones

Se concluye que los alumnos que se pronostican morosos son los que no cuentan con asistencia de la Universidad, conjuntamente, cuentan con padres con ingresos menores a 1200 Nuevos Soles y la particularidad de cancelación elegida es 5 cuotas, también de mostrar un contexto laboral emancipado. Por el contrario, los alumnos pronosticados como No Morosos dependen claramente de la entrada de los padres ya que los que tiene padres con ingresos mayores a 2000 soles poseen un porcentaje bajo de riesgo de aproximadamente un 75%. El modelo Back Propagación basado en redes neuronales permite la caracterización de los distintos perfiles según la conducta del alumno permitiendo a la administración de hacienda alumnos poseer un superior control y una dirección más eficaz del peligro crediticio, de esta manera se pueda instituir estrategias correspondientes para reducir el incumplimiento en el desembolso de sus armadas en la Universidad Peruana Unión

**Punto de vista:**

El paper anterior nos beneficia en la investigación de una muy buena manera, ya que nos da ejemplo claro de la forma de usar la metodología CRISP-DM la cual será utilizada en el desarrollo de la tesis, muy aparte de darnos un nueva forma de representarlo usando un modelo de back-propagation, mostrándonos así un nuevo énfasis de desarrollo del modelo predictivo, usando en este caso las redes neuronales como medio de identificación de los morosos en la universidad donde se ha de aplicar, nos da un nueva visión, del desarrollo ya que se ajusta al tema de tesis a desarrollar.

## 2.2 MARCO TEÓRICO

### Crisp – DM (Rodriguez, 2014)

Para implementar una tecnología en un negocio, se requiere de una sistemática. La colectividad de las consultoras especializadas en alguna know-how cuentan, con por lo menos, una sistemática, según los tipos de proyectos que aborden. Estos métodos son definidos a partir de sus experiencias y tomando lo superior de los procedimientos crecidamente exitosos o populares. Referir con una metodología, se ha transformado tan trascendental y preciso como la esquela de exposición de las empresas. Para los diferentes tipos de tecnologías, hay varias metodologías, algunas están publicadas en Internet. Para el asunto de proyectos de implementación de explotación de datos, hay una en particular; CRISP-DM, en sus primeros años de propaganda tenía apoyos de empresas privadas y organismos públicos, pero poco a poco ha ido perdiendo uno que otro "Project Partner". Desconocemos la causa de esta supuesta perdida de sostén, pero estamos seguros que no corresponde a la imperfección de calidad o efectividad del método, porque ha sido adoptado por otros organismos y empresas.

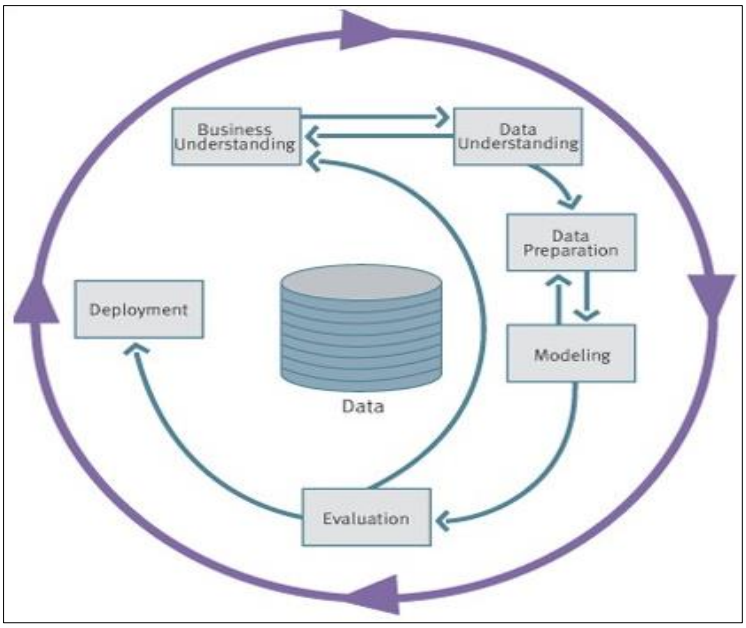


*Figura 7.* Grafica metodología CRISP-DM
Adaptado de "Grafica CRISP-DM" por (Rodríguez, 2014)

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)

Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)

Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio)

- Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas) Comprensión de los datos (Familiarizarse con los datos teniendo presente los objetivos del negocio)
- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

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

- Selección de los datos
- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

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

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

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

- Evaluación de resultados
- Revisar el proceso
- Establecimiento de los siguientes pasos o acciones

**Despliegue (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización) (UTM, 2012)**

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Generación de informe final

**Metodología CRISP-DM (Goicochea, 2012)**

La sistemática CRISP-DM consta de cuatro niveles de ensimismamiento, organizados de grafía jerárquica en tareas que van desde el horizonte más corriente incluso los casos crecidamente específicos (ver Figura 7). A nivel más ordinario, el sumario está organizado en seis fases (ver Figura 8), estando cada período a su vez estructurada en varias tareas generales de segundo horizonte o sub fases. Las tareas generales se proyectan a tareas específicas, en que se describen las acciones que deben ser desarrolladas para situaciones específicas. Asimismo, si en el segundo horizonte se tiene la tarea corriente "limpieza de datos", en el tercer nivel se dicen las tareas que tienen que desarrollarse para un asunto específico, como, por ejemplo, "limpieza de datos numéricos", o "limpieza

de datos categóricos". El cuarto horizonte, recoge el conjunto de acciones, decisiones y resultados sobre el plan de aprovechamiento de información especifico.

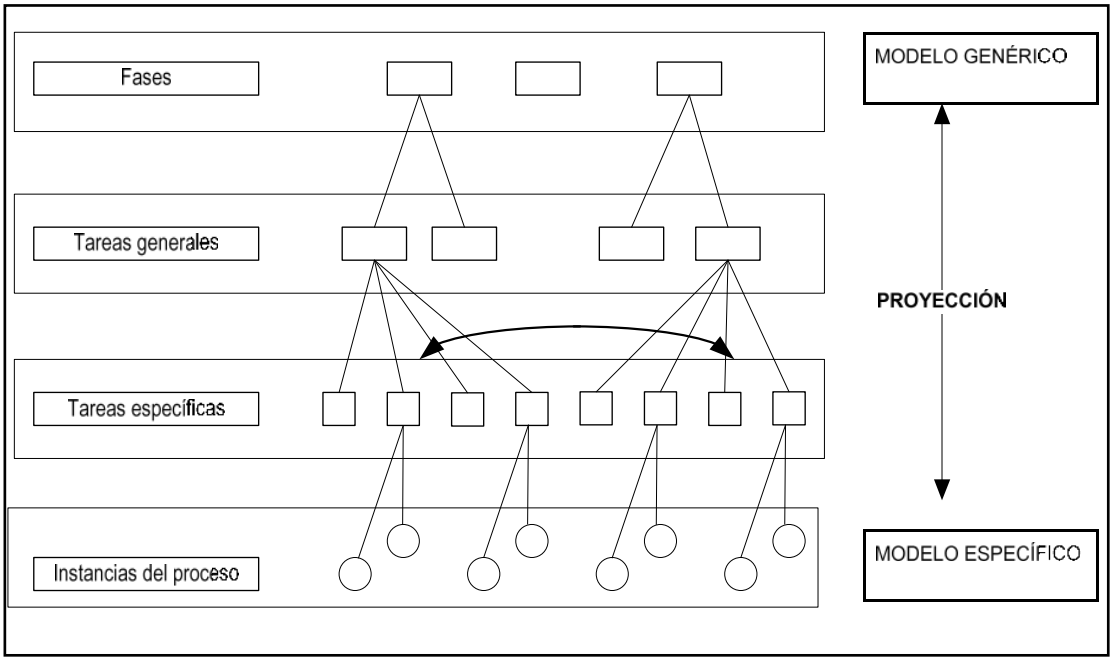


*Figura 8.* Esquema de los cuatro niveles de abstracción de la metodología CRISP-DM
Adaptado de "Esquema de metodología CRISP-DM" por (Goicochea, 2012)

La metodología CRISP-DM proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de Explotación de Información: el modelo de referencia y la guía del usuario.

El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de Explotación de Información en general. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de Explotación de Datos específicos, proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase. La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Explotación de Información en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto (Figura 8).

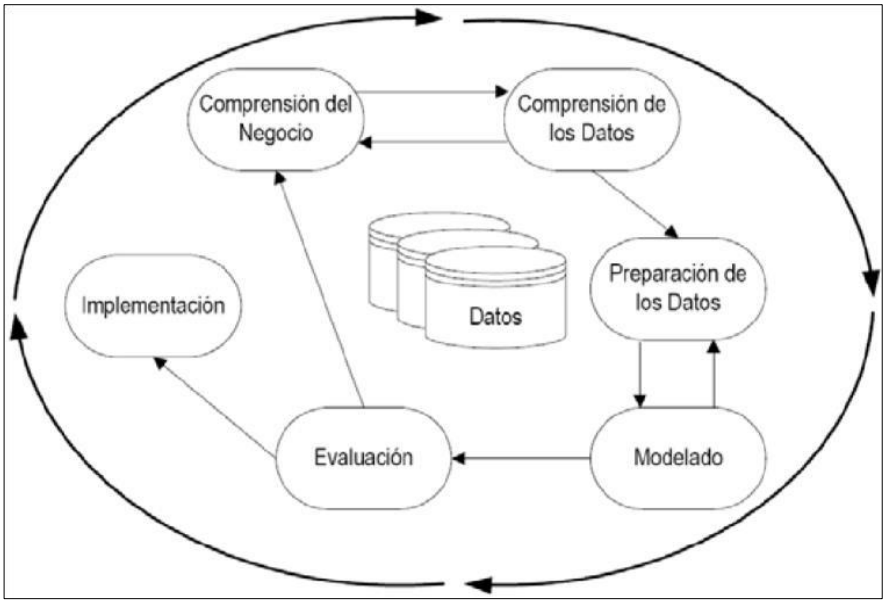


*Figura 9.* Fases del proceso de modelado metodología CRISP-DM.

Adaptado de "Fases de proceso de la metodología CRISP-DM" por (Goicochea, 2012)

Las flechas indican las relaciones más y más habituales entre las fases, no obstante, se pueden instituir relaciones entre fases cualesquiera. El círculo externo simboliza el hábitat cíclico del sumario de modelado. En la Figura 9, se detallan las fases que componen a la sistemática CRISP-DM y en la tabla 12, se detalla cómo se componen vuelta una de ellas.

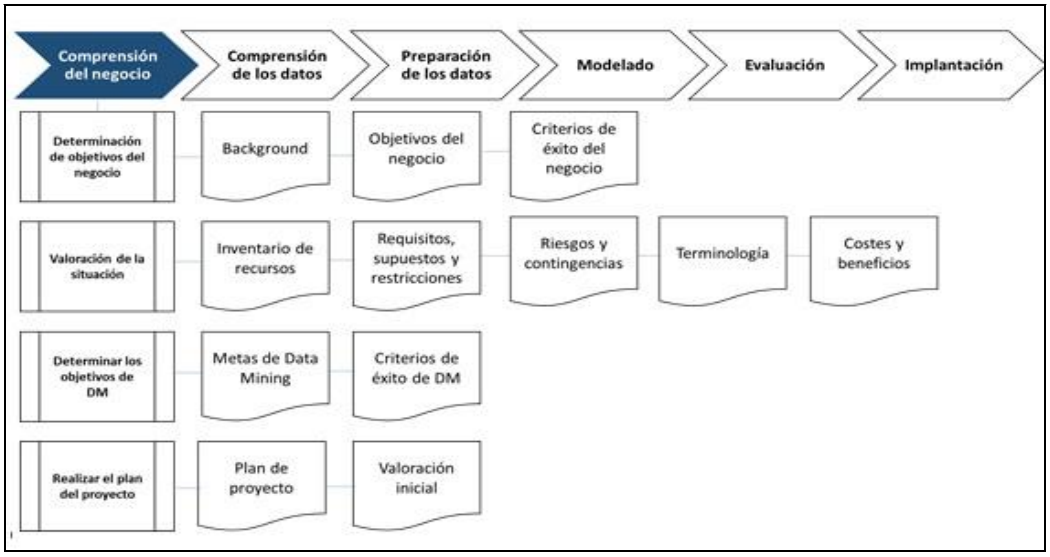


*Figura 10.* Fases de la metodología CRISP-DM. (Goicochea, 2012)

Adaptado de "Fases de la metodología CRISP-DM" por (Goicochea, 2012)

La primera etapa de estudio del inconveniente incluye la agudeza de los objetivos y requerimientos del plan desde una representación institucional, con el desenlace de convertirlos en objetivos técnicos y en una planificación. La segunda etapa de estudio de datos comprende la cosecha inicial de datos, en disposición a que sea potencial instituir una primera relación con la dificultad, identificando la aptitud de los datos y estableciendo las relaciones más evidentes que permitan instituir las primeras suposición.

Una sucesión realizada el examen de datos, la sistemática establece que se proceda al apresto de los datos, de tal grafía que puedan ser tratados por las técnicas de modelado. La gestación de datos incluye las tareas generales de opción de datos a los que se va a emplear la habilidad de relieve (variables y muestras), lavado de los datos, reproducción de variables adicionales, composición de diferentes orígenes de datos y cambios de formato. El período de gestación de los datos se encuentra estrechamente relacionada con la etapa de relieve, pareja que en función de la pericia de formato que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Por lo tanto, las fases de preparativo y modelado interactúan de grafía sistemática.

Antes de proceder al modelado de los datos se debe de establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos.

Una vez realizadas estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos. En la fase de evaluación, se evalúa el modelo, no desde el punto de vista de los datos, sino desde el cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la aplicación del modelo

Tabla 11
*Tareas de cada fase de la metodología CRISP-DM*

| | TAREAS COMPONENTES | ACTIVIDADES ASOCIADAS |
|---|---|---|
| Comprensión del negocio | Determinar los objetivos del negocio | • Background<br>• Objetivos del negocio<br>• Criterios de éxito del negocio |
| | Evaluar la situación | • Inventarios de recursos<br>• Requisitos y requerimientos<br>• Riesgos y contingencias<br>• Terminología<br>• Costos y beneficios |
| | Determinar objetivos del proyecto de Explotación de Información | • Las metas del Proyecto de<br>• Explotación de Información<br>• Criterios de éxito del Proyecto de Explotación de Información |
| | Realizar el Plan del Proyecto | • Plan de proyecto<br>• Valoración inicial de herramientas |
| Comprensión de los datos | Recolectar los datos Iniciales | • Reporte de recolección de datos iniciales |
| | Descubrir datos | • Reporte de descripción de los datos |
| | Explorar los datos | • Reporte de exploración de datos |
| | Verificar la calidad de datos | • Reporte de calidad de datos |
| | Caracterizar el conjunto de datos | • Conjunto de Datos<br>• Descripción del Conjunto de Datos |
| Preparación de los datos | Seleccionar los datos | • Inclusión / exclusión de datos |
| | Limpiar los datos | • Reporte de calidad de datos limpios |
| | Estructurar los datos | • Derivación de atributos<br>• Generación de registros |
| | Integrar los datos | • Unificación de datos |
| | Caracterizar el formato de los datos | • Reporte de calidad de los datos |
| | Seleccionar una técnica de modelado | • La técnica modelada<br>• Supuestos del modelo |
| | Generar el plan de pruebas | • Plan de pruebas |

| | | |
|---|---|---|
| Modelado | Construir el modelo | • Configuración de parámetros<br>• Modelo<br>• Descripción del modelo |
| | Evaluar el modelo | • Evaluar el modelo<br>• Revisación de la configuración de parámetros |
| | Evaluar Resultado | • Valoración de resultados mineros con respecto al éxito del negocio<br>• Modelos aprobados |
| Evaluación | Revisar | • Revisión del proceso |
| | Determinar próximos pasos | • Listar posibles acciones |
| Implementación | Realizar el plan de implementación | • Plan de Implementación |
| | Realizar el plan de monitoreo y mantenimiento | • Plan de monitoreo y mantenimiento |
| | Realizar el informe final | • Informe final<br>• Presentación Final |
| | Realizar la revisión del proyecto | • Documentación de la experiencia |

## Conclusión

Normalmente los proyectos de aprovechamiento de información no terminan en la institución de modelo, sino que se deben justificar los resultados de modo comprensible en orden a lograr un incremento del discernimiento. Conjuntamente en la etapa de aprovechamiento se debe de afirmar el sustento de la diligencia y la propagación de los resultados.

## Modelado en CRISP-DM (Pete, 2011)

El modelo y cumplimiento de la explotación de información toma parte en la etapa de formato (Fase 4) de esta sistemática. Se prueban suposiciones especificas y se ejecutan métodos de revelación automatizadas, se interpretan los resultados de estudio realizados en esta etapa en el tejido de las preguntas del oficio originales. En la figura 10 se detalla cómo está compuesta la etapa. Las tareas que se realizan son:

elegir una habilidad del modelado, crear el procedimiento de pruebas, edificar el modelo y tasar el modelo.

**Seleccionar una técnica de modelado**

Como primer paso a ejecutar, se selecciona la técnica de modelado a utilizar. Considerando que ya, posiblemente, se seleccionó una herramienta de negocio, esta tarea se refiere a la técnica de modelado específica, por ejemplo, árboles de decisión, reglas de decisión, redes neuronales, etc. Si se considera necesario aplicar múltiples técnicas, se debe realizar esta tarea, para cada una de las técnicas, separadamente. No se debe olvidar que no todas las herramientas y técnicas son aplicadas en cada tarea. Para determinados tipos de problemas, algunas técnicas son las apropiadas. A continuación, se detalla la relación existente

**Descripción de datos:**

Consiste en la descripción de las características de los datos, típicamente en formas elementales y de agregación. Esto da al usuario una muestra de la estructura de los datos. En los proyectos de explotación de información la descripción de datos y la sumarización son un sub objetivo del proceso, típicamente en tempranas etapas. La exploración y el análisis inicial de los datos pueden ayudar a entender la naturaleza de ellos y encontrar hipótesis potenciales de información oculta. La estadística descriptiva y las técnicas de visualización proveen una primera visión de los datos.

**Segmentación:**

Tiene por ecuánime la ausencia de los datos en subgrupos o clases interesantes. Todos los elementos del subgrupo deben poseer características comunes. El examen de las conjeturas de los subgrupos es notable para los cuestionamientos bases del negocio referente a la base de la escapatoria de la imagen de los datos y la sumarización. Las técnicas apropiadas para segmentar son: Técnicas de clustering, redes neuronales y visualización.

**Descripción de conceptos:**

Tiene por objetivo entender la descripción de los conceptos o clases. La descripción de conceptos tiene relación con la segmentación y con la clasificación. La segmentación puede conducir a una enumeración de objetos que pertenecen a un

concepto o la clase sin una descripción comprensible. Típicamente hay segmentación antes de que la descripción de concepto sea realizada. Algunas técnicas, por ejemplo, clustering, realizan la segmentación y la descripción de conceptos al mismo tiempo. La descripción de conceptos puede ser usada también como una clasificación de propósitos. Las técnicas apropiadas son: Métodos de reglas de inducción, clustering conceptuales

**Clasificación:**

La clasificación asume que hay un vinculado de objetos (caracterizados por algunos atributos) en los cuales hay diferentes clases. El rtulo de la clases es de importe prudente y se conoce cadencia de objeto.
El objetivo es conseguir modelos de codificación (clasificadores) los cuales determinen educadamente la clase ante objetos no previstos precedentemente. Los modelos de clasificación referente a todo son usados para el formado predictivo. Los rótulos de clases avanzados pueden ser definidos por el beneficiario o derivados de la segmentación. Las técnicas apropiadas para este tipo de dificultad son: Análisis de la discriminante, métodos de incitación de reglas, árboles de fallo, árboles de noviciado, redes neuronales, vecino más cercano, casos basados en lógica y algoritmos genéticos.

**Análisis de dependencias:**

Pueden ser usadas como valores de predicción de un dato, teniendo información de los otros datos. A través de las dependencias puede usarse un modelo predictivo. Las asociaciones son una clase especial de dependencias, las asociaciones describen afinidad entre los ítems. El análisis de dependencias tiene relaciones con la clasificación y la predicción, donde las dependencias están implícitamente usadas para la formulación de modelos predictivos. Las técnicas aplicadas son: Análisis de correlación, análisis de regresión, reglas de asociación, redes bayesianas, programación lógica inductiva, técnica de visualización.

Entre las herramientas y técnicas hay “requisitos políticos” y otras restricciones, que limitan la elección de herramientas. Puede ser que simplemente un instrumento o pericia esté aprovechable para remediar el inconveniente, en cuyo asunto puede ocurrir que el utensilio no sea la mejor para resolverlo. Los reportes de las tareas realizadas mientras esta sub-fase son:

**La técnica modelada:**

Descripción de las técnicas de modelado que se utilizarán. Supuestos de modelado: Muchas técnicas generan supuestos específicos en los datos, por ejemplo, todos los atributos tienen una distribución uniforme, o no existen valores perdidos. Todos estos supuestos deben ser registrados

**Generar el Plan de Pruebas**

Antes de generar el modelo se debe generar un procedimiento o mecanismo para probar la calidad y validez del modelo. El reporte de la tarea realizada durante esta sub fase es:

Plan de Pruebas: Se debe describir el plan de pruebas y los modelos. Un componente principal del plan de pruebas es cómo dividir el conjunto de datos disponible en datos de entrenamiento y datos de validación.

**Construir el modelo**

Se debe hacer el arma de modelado con el ligado de datos dispuesto para fundar uno o más modelos. Los reportes de las tareas realizadas mientras esta sub etapa son: Disposición de en general, la mayoría de las herramientas de modelado proveen un parámetro: conjunto de parámetros de arreglo a concordar. Se debe listar el conjunto de parámetros y los valores escogidos para los mismos.

**Modelo: Describir los modelos reales generados por la herramienta.**

Descripción del Modelo: Se describe el modelo resultante, mediante un informe que detalle la interpretación de los modelos y documente cualquier dificultad encontrada con su significado.

**Conclusión**

Los dataminers deben descifrar los modelos según su potestad de discernimiento, los datos, el criterio de triunfo y el plan de pruebas determinado. Esta labor interfiere con el período subsiguiente, considerando que los datos que se "Explotan" a cordura del dataminer definen el éxito de la diligencia de relieve y las técnicas de revelación. El dataminer comunica a los analistas del ejercicio y expertos en el mando de la diligencia los resultados obtenidos, para discutir con éstos los resultados de la utilización de datos en el domino del ejercicio; intentando alinear los

datos de la organización a los modelos y analizar los modelos según los criterios de evaluación. En la mayoría de los proyectos, el dataminer aplica la misma técnica más de una vez o intenta generar los resultados con técnicas alternativas. Los reportes de las tareas realizadas durante esta sub fase son:

Evaluación del modelo: Se deben resumir los resultados de la tarea, detallando la calidad de los documentos generados.

Revisión de parámetros Según las valoraciones, se deben revisar las configuraciones de los de configuración: parámetros para las próximas corridas del modelo. Se debe, también, iterar el modelo construido y la configuración de los parámetros hasta encontrar el mejor modelo, documentando todas las revisiones y valoraciones.

**Árbol de decisión (Nivel, 2014)**

**DEFINICION:**

Árboles de Decisión. Técnica que permite analizar decisiones secuenciales basada en el uso de resultados y probabilidades asociadas.

Los árboles de decisión se pueden usar para generar sistemas expertos, búsquedas binarias y árboles de juegos, los cuales serán explicados posteriormente.

Las ventajas de un árbol de decisión son:

Resume los ejemplos de partida, permitiendo la clasificación de nuevos casos siempre y cuando no existan modificaciones sustanciales en las condiciones bajo las cuales se generaron los ejemplos que sirvieron para su construcción.

Facilita la interpretación de la decisión adoptada.

Proporciona un alto grado de comprensión del conocimiento utilizado en la toma de decisiones.

Explica el comportamiento respecto a una determinada tarea de decisión.

Reduce el número de variables independientes.

Es una magnifica herramienta para el control de la gestión empresarial.

Los árboles de decisión se utilizan en cualquier proceso que implique toma de decisiones, ejemplos de estos procesos son:

- Búsqueda binaria.
- Sistemas expertos.
- Árboles de juego

Los árboles de decisión generalmente son binarios, es decir que cuentan con dos opciones, aunque esto no significa que no puedan existir árboles de tres o más opciones.

## BÚSQUEDA BINARIA

Búsqueda binaria es el método en el cual la búsqueda partiendo al árbol en dos partes

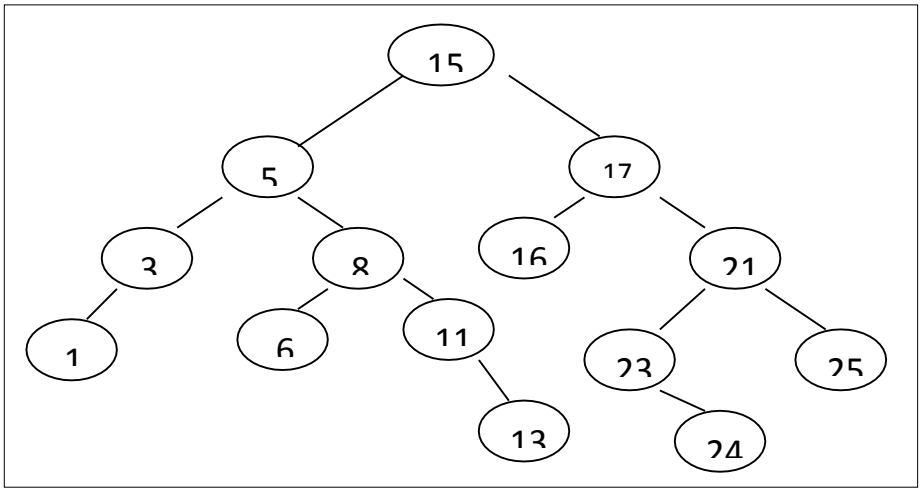


*Figura 11*. Búsqueda Binaria adaptado de "Búsqueda Binaria" de (Nivel, 2014)

Supongamos que deseas buscar un numero x en el árbol.

Comparamos si el dígito que estamos buscando es semejante a la cepa, si es igual se devuelve la raíz y se termina la investigación. Si no es igual se compara reiteradamente el digito para estar al tanto si es mayor o menor que la raíz con lo que se despreciaría la mitad del árbol volviendo la investigación más rápida. Si es menor recorremos la indagación hacia el lado corvo hasta hallar el siguiente elemento de árbol, el cual volvemos a cotejar como lo hicimos con la raíz. Si es mayor se realiza la investigación hacia el paraje derecho del árbol, el cual lo tomamos como si fuera una raíz y comparamos de la misma forma que la primera raíz.

Los pasos anteriores se realizan hasta encontrar el elemento buscado o llegar a NULL que nos indicaría que el elemento no se encuentra en el árbol.

## ÁRBOLES DE JUEGO

Los árboles de juego son una aplicación de los árboles de decisión. Tomemos por ejemplo el conocido juego del gato y consideremos una función evalúa que acepta una posición del tablero y nos devuelve un valor numérico (entre más grande es este valor, más "buena" es esta posición). Un ejemplo de la implantación de esta función es considerando el número de renglones, columnas y diagonales restantes abiertas para un jugador menos el número de las mismas para su oponente, por ejemplo, la sig. Posición en un juego y sus posibles continuaciones:

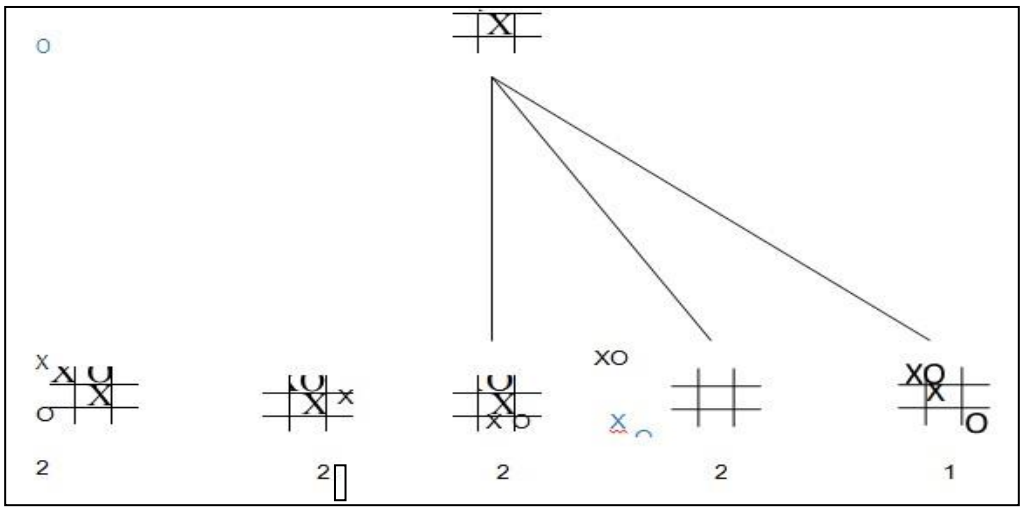


*Figura 12*. Grafica de árbol de juego adaptado de "Grafica de árbol de juego" por (Nivel, 2014)

Dada una posición del tablero, el mejor movimiento siguiente está determinado por la consideración de todos los movimientos posibles y las posiciones resultantes. Tal análisis no conduce sin embargo al mejor movimiento, como se ve en el ejemplo

Primero cuando las cuatro primeras posibilidades dan indiviso el propio importante de valoración, sin embargo, el pedazo posición es sin duda mejor, por lo que se debe optimizar esta función. Ahora se introduce la contingencia de prever varios movimientos. En aquel tiempo la función se mejorará en gran medida, se inicia con cualquier enfoque y se determinan todos los posibles movimientos en un árbol hasta un animoso nivel de conjetura. Este árbol se conoce como "árbol de juego" cuya hondura es igual a la hondura de dicho árbol. El sig. Árbol de esparcimiento para la posición para la posición inicial del felino y un horizonte de previsión se muestra a continuación

Designamos el turno del jugador 1 como +, y el turno del jugador 2 como -, es claro que como el árbol empieza con el turno de +, entonces el árbol estará evaluado de acuerdo a la conveniencia de +. De acuerdo al árbol anterior el mejor primer turno para + será la cruz en el centro por lo que el jugador decidirá hacer este movimiento, en esta fase se ve que el turno que sigue es de -, - deberá seleccionar la jugada que tenga el menor valor, pues esta será la que perjudique más a + y convendrá a -.

Así es como funciona un árbol de juego que es una diligencia de un árbol de arbitraje, puesto que se genera el árbol de acuerdo al nivel de conjetura y cada jugador va decidiendo que se apuesta le conviene más de acuerdo a la estimación de una determinada posición

**Conclusión**

Los árboles de fallo se usan en los sistemas expertos ya que son más precisos que el hombre para poder desarrollar un diagnostico con relación a algo, ya que el hombre puede dejar franquear sin querer un detalle, en cambio la maquina mediante un método experto con un árbol de arbitraje puede dar un resultado cabal. Una deficiencia de este es que puede alcanzar a ser más lento pues analiza todas las posibilidades, dificultad esto a su vez es lo que lo vuelve más exacto que al hombre. A continuidad, se presenta un modelo de un sistema experto y de cómo puede llegar a determinar que se emplee un fármaco X en una persona con presión circulatorio. Algoritmos de minerías de datos (Inacap, 2014).

**1.1 Definición.**

Un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de minería de datos a partir de los datos.

*Pasos para crear un modelo:*

- El algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias.
- El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos.

A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

**Algoritmos de Clasificación:**

Son aquellos que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos.

**Arboles de decisión.**

Un árbol de decisión predice un resultado y describe cómo afectan a estos diversos criterios.

**Cuando elegir un algoritmo de Clasificación:**

Marcar los clientes de una lista de posibles compradores como clientes con buenas o malas perspectivas.

Calcular la probabilidad de que un servidor genere un error en los próximos 6 meses.

Clasificar la evolución de los pacientes y explorar los factores relacionados.

**Ejemplo:**

El departamento de marketing de la empresa desea identificar las características de los clientes antiguos que podrían indicar si es probable que realicen alguna compra en el futuro. La base de datos almacena información demográfica que describe a los clientes antiguos. Mediante el algoritmo de árboles de decisión que analiza esta información, el departamento puede generar un modelo que predice si un determinado cliente va a comprar productos, basándose en el estado de las columnas conocidas sobre ese cliente, como la demografía o los patrones de compra anteriores.

**Funcionamiento:**

El algoritmo de árbol de decisión genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

**Predecir columnas discretas**

La forma en que el algoritmo de árboles de decisión genera un árbol para una columna de predicción discreta puede mostrarse mediante un histograma. El siguiente diagrama muestra un histograma que traza una columna de predicción, según una columna de entrada, Edad.

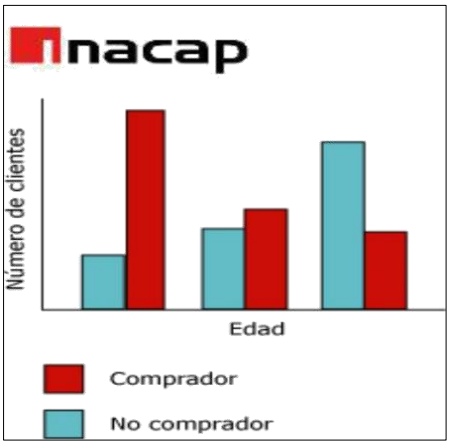


*Figura 13.* Grafica Cliente – No cliente
Adaptado de "Grafica Cliente-No cliente"
por (Inacap, 2014)

La correlación que aparece en el diagrama hará que el algoritmo de árboles de decisión de creer un nuevo nodo en el modelo.

Histograma: Gráfico de la representación de distribuciones de frecuencias, en el que se emplean rectángulos dentro de unas coordenadas.

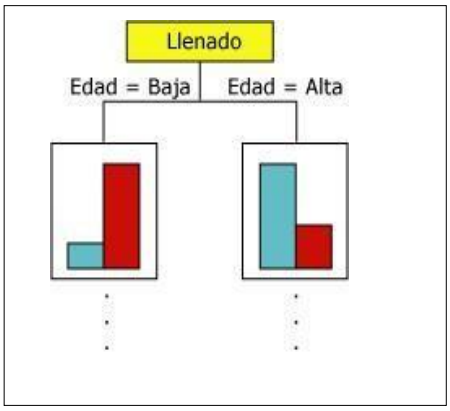


*Figura 14*. Grafica Cliente – No cliente
Adaptado de "Grafica Cliente-No cliente"
por (Inacap, 2014)

A medida que el algoritmo agrega nuevos nodos a un modelo, se forma una estructura en árbol. El nodo superior del árbol describe el desglose de la columna de predicción para la población global de clientes. A medida que el modelo crece, el algoritmo considera todas las columnas.

### 1.2 Datos requeridos para los modelos de árboles de decisión

Cuando prepare los datos para su uso en un modelo de árboles de decisión, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos se utilizan.

Los requisitos para un modelo de árboles de decisión son los siguientes:

Una 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.

Una columna de predicción: Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.

Columnas de entrada: Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

Para obtener información más detallada sobre los tipos de contenido y los tipos de datos admitidos en los modelos de árboles de decisión, vea la sección Requisitos de Referencia técnica del algoritmo de árboles de decisión de.

**Ventajas:**

Evitar sobreajuste de datos.

- Determinación de la profundidad de crecimiento del árbol.
- Reducción de errores en la poda.
- Condicionamiento de la poda
- Manejo de atributos continuos.
- Mejora en la eficiencia computacional.

- Simpleza y legibilidad para la interpretación de los resultados.

**Algoritmo C4.5 - J45.**

Es una extensión del algoritmo ID3, el cual genera un árbol de decisión a partir de datos mediante las participaciones recursivas utilizando una estrategia de profundidad.

Este algoritmo utiliza la heurística para establecer una proporción de ganancia.na prueba con n número de resultados, siendo n el número de valores posibles que puede tomar el atributo.

Se consideran todas las pruebas posibles para dividir el conjunto de datos y selecciona la prueba que la haya generado la mayor ganancia de información. Para cada atributo discreto, se considera

**1.3 Pseudocódigo C4.5:**

- Conjunto de atributos no clasificados.
- Atributo clasificador.
- Conjunto de entrenamiento, devuelve un árbol de decisión.
- Comienzo

**Si S está vacío.**

Devolver un único nodo con valor FALLA; para formar el nodo raíz

Si Todos los registros de S tienen el mismo valor para el atributo clasificador.

Devolver un único nodo con dicho valor; un único nodo para todos. Si R está vacío.

Devolver un único nodo con el valor más frecuente del atributo clasificador en los registros de S [existirán errores los cuales no estarán bien clasificados]

**Si R no está vacío.**

Devolver atributo con mayor proporción de ganancia entre los atributos de R; Sean (dj - > j-1, 2,…. m) los valores del atributo D:

Sean (dj -> j-1, 2,…, m) los valores del atributo de S correspondientes a los valores de dj respectivamente

Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d1, d2,…, dm, que

Van respectivamente a los árboles. (Inacap, 2014)

### 2.2.1 Bases Teóricas

La Minería de Datos (Data Mining) es el proceso de extraer información no trivial y potencialmente útil a partir de grandes conjuntos de datos disponibles en las ciencias experimentales (registros históricos de observaciones, reanálisis, simulaciones de GCMs, etc.), proporcionando información en un formato legible que puede ser usada para resolver problemas de diagnosis, clasificación o predicción. Tradicionalmente, este tipo de problemas se resolvían de forma manual aplicando técnicas estadísticas clásicas, pero el incremento del volumen de los datos ha motivado el estudio de técnicas de análisis automáticas que usan herramientas más complejas. Por lo tanto, la Minería de datos identifica tendencias en los datos que van más allá de un análisis simple. Técnicas modernas de Minería de datos (reglas de asociación, árboles de decisión, modelos de mezcla de Gaussianas, algoritmos de regresión, redes neuronales, máquinas de vectores soporte, Redes Bayesianas, etc.) se utilizan en ámbitos muy diferentes para resolver problemas de asociación, clasificación, segmentación y predicción. (Gutiérrez, 2013)

Entre los diferentes algoritmos de Minería de datos, los modelos gráficos probabilísticos (en particular las Redes Bayesianas) constituyen una metodología elegante y potente basada en la probabilidad y la estadística que permite construir modelos de probabilidad conjunta manejables que representan las dependencias relevantes entre un conjunto de variables (cientos de variables en aplicaciones prácticas). Los modelos resultantes permiten realizar inferencia probabilística de una manera eficiente. Por ejemplo, una Red Bayesiana podría representar las relaciones probabilísticas entre campos sinópticos de larga escala y registros de observaciones locales, proporcionando una nueva metodología de downscaling probabilístico: p. ej. permite calcular P (observación predicción de larga escala). Por ejemplo, en la siguiente figura los puntos rojos representan nodos de la rejilla de un GCM, mientras

que los puntos azules corresponden a estaciones con registros de observaciones (los enlaces muestran las dependencias importantes aprendidas de forma automática a partir de los datos). (Gutiérrez, 2013)

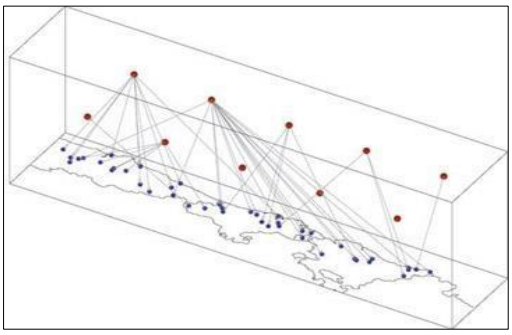

*Figura 15*. Grafica Red bayesiana adaptado de "Grafica red bayesiana" por (Gutiérrez, 2013)

Formalmente, una Red Bayesiana es un grafo dirigido a cíclico cuyos nodos representan variables y los arcos que los unen codifican dependencias condicionales entre las variables. El grafo proporciona una forma intuitiva de describir las dependencias del modelo y define una factorización sencilla de la distribución de probabilidad conjunta consiguiendo un modelo manejable que es compatible con las dependencias codificadas. Existen algoritmos eficientes para aprender modelos gráficos probabilísticos a partir de datos, permitiendo así la aplicación automática de esta metodología en problemas complejos. Las Redes Bayesianas que modelizan secuencias de variables (por ejemplo, series temporales de observaciones) se denominan Redes Bayesianas Dinámicas. Una generalización de las Redes Bayesianas que permiten representar y resolver problemas de decisión con incertidumbre son los Diagramas de Influencia.

Por otra parte, las redes neuronales son modelos no lineales, inspirados en el funcionamiento del cerebro, que fueron diseñados para resolver una gran variedad de problemas. Los perceptrones multi-capa son algoritmos de regresión que construyen un modelo determinista y=f(x), relacionando un conjunto de predictores, x, y predictandos, y (figura inferior izquierda). Las redes auto-organizativas (SOM) son redes competitivas diseñadas para problemas de agrupación (clustering) y visualización (figura inferior derecha). (Gutiérrez, 2013)

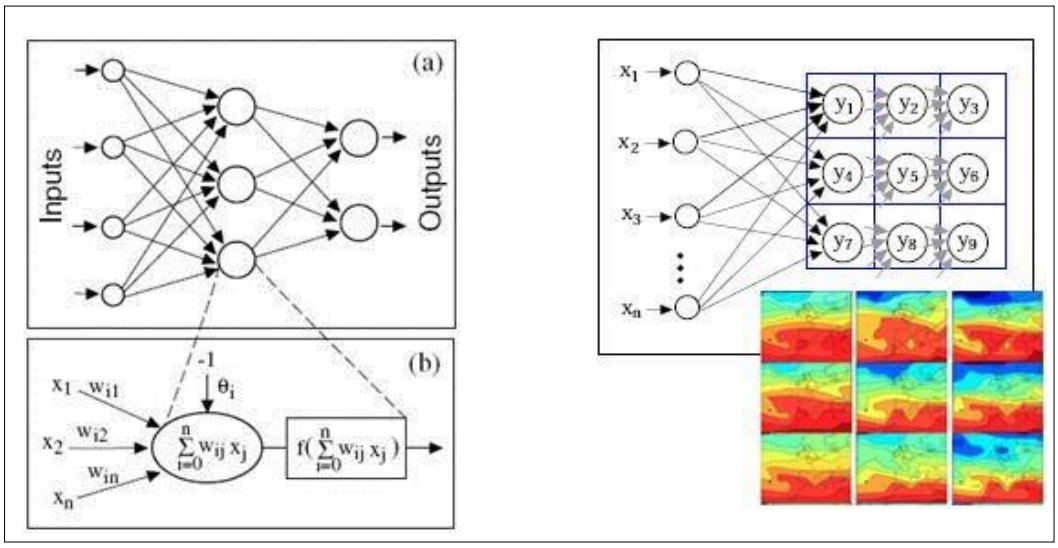


*Figura 16.* Grafica Red bayesiana adaptado de "Grafica red bayesiana" por (Gutiérrez, 2013)

**Conclusiones**

La importancia y uso creciente de los nuevos modelos, representa la inmersión en un nuevo mundo en el cual la incertidumbre no constituye un impedimento para un eficaz tratamiento en la toma de decisiones y evaluación de procesos.

La combinación de un adecuado sistema de inferencia con el conocimiento adquirido durante la elaboración de un modelo bayesiano por parte de los expertos, constituye una gran fortaleza para cualquier análisis, puesto que existen factores externos al sistema que proporcionan información adicional para la evaluación del requerimiento.

La teoría de la decisión bayesiana es ideal en aplicación para la solución de problemas de comercialización, teniendo en cuenta los parámetros de la incertidumbre; donde la incertidumbre debe tener en cuenta la toma de decisiones como una acción que establece el valor de diversas variables en el entorno de mercado que enfrenta el consumidor y las compañías.

El método bayesiano puede ser apropiado si los gerentes a cargo de la toma de Decisiones están dispuestos a utilizar un modelo que tenga en cuenta sus conocimientos y experiencia.

**METODOLOGÍA SEMMA (Pablo, 2012)**

El programa SAS Enterprise Miner utiliza la metodología SEMMA. Cada una de las iniciales hace referencia a cada una de las fases de un proyecto de minería de datos, que a su vez contienen diferentes nodos que el analista puede escoger en función de qué modelo (descriptivo o predictivo) quiera llevar a cabo.

SAMPLE: es la primera etapa del proyecto. En ella preparamos los datos para su posterior exploración. En esta etapa es común la utilización del nodo de partición (especialmente si quieren realizarse árboles de decisión o redes neuronales). Normalmente se suele utilizar un porcentaje de 70 para la muestra de entrenamiento y uno de 30 para la validación

EXPLORE: se trata de la exploración de los datos. Es una de las partes más trabajosas, pero también la más bonita. Tenemos un nodo que nos ayudan a explorar gráficamente los datos, otro de selección de variables que nos ayuda a eliminar aquellos inputs que no tienen relación con la variable objetivo, incluso podemos hacer un "clustering" o una segmentación

MODIFY: cuando llegamos a esta parte ya empezamos a hablar en serio. Aquí nos centramos en la selección y transformación de variables y datos que servirán para la construcción de los modelos. Entre otras tareas a realizar destacan: la reducción de dimensión, imputación de valores "missing", "outliers", etc

MODEL: ha llegado la hora de escoger los modelos. La elección del modelo va a depender esencialmente de los datos que tenemos y del tipo de variables que tenemos y de obtener modelos fácilmente entendibles. Podemos escoger regresión, regresión logística, árboles de decisión, análisis factorial discriminante, redes neuronales... Podemos aplicar más de uno a la vez, y luego comparar los resultados obtenidos

ASSESS: después de todo el trabajo realizado llega el momento de comparar los modelos. Lo más sencillo es utilizar el análisis del diagrama ROC. La curva ROC es útil para comparar el comportamiento global de un modelo. El gráfico ROC enfrenta dos variables: la sensibilidad y la especificidad. Lo ideal es que ambas categorías sean altas.

## Conclusiones

El desarrollo de las bases de datos y los sistemas de computación han generado gran cantidad de información que sólo puede ser justificada si se utiliza como fuente de información para mejorar el proceso en el que es generada. Sin embargo dada la novedad del sector y la característica de I+D del proceso de análisis, éste no se realiza de forma suficientemente estructurada, por lo que se producen grandes errores en las estimaciones de coste y plazo en este tipo de proyectos.

La utilización de una metodología estructurada y organizada presenta las siguientes Ventajas para la realización de proyectos de Data Mining:

- Facilita la realización de nuevos proyectos de Data Mining con características similares
- Facilita la planificación y dirección del proyecto
- Permite realizar un mejor seguimiento del proyecto

## Tabla de comparación de metodologías

Tabla 12
*Comparación de metodologías*

| | **SEMMA** | **CRISP-DM** |
|---|---|---|
| **Permite elección libre de las herramientas** | NO | 12 |
| **Cantidad de fases** | 5 | 6 |
| **Todas las fases pueden relacionarse** | NO | SI |
| **Considera los motivos del proyecto** | NO | NO |
| **Considera la naturaleza del interés de la parte** | NO | NO |
| **Considera otros aspectos no técnicos** | NO | SI |
| **Identificar claramente las variables** | NO | NO |
| **Esta detallada paso a paso cada etapa del método** | NO | NO |
| **Identificar problemas inteligencia de negocio** | NO | SI |
| **Identificar técnicas de explotación de Información** | Parcialmente | NO |
| **Identificar relaciones entre las TEI y los PIN** | SI | SI |
| **Identificar procesos a explotar** | Parcialmente | NO |
| **Identificar procesos de explotación de información** | Parcialmente | NO |

# CAPÍTULO III
# DESARROLLO DEL MODELO PREDICTIVO

## 3.1 ESTUDIO DE FACTIBILIDAD

### 3.1.1 Factibilidad Técnica

Disponemos de la plataforma necesaria para poder demostrar nuestro estudio, apoyándonos en las metodologías que vamos a aplicar, muy a parte del hardware, que nos ayudara a hacer más rápido los procesos que vayamos a desempeñar

Tabla 13
*Factibilidad Técnica*

| **Técnicos** | |
|---|---|
| **Computadora** | 2000 |
| **Impresora** | 180 |
| **Paquete Office** | 219 |

### 3.1.2 Factibilidad Operativa

Al desarrollarse el modelo predictivo se respalda por valores que se muestran en el estudio será usado para mejorar los procesos de la Universidad donde se hará. El modelo predictivo es necesario para poder mejorar el estudio cualitativo de las características de los posibles alumnos deudores. Al haber personas con experiencia empírica se dispone del conocimiento de ellas para que se pueda alimentar de mejor forma el modelo predictivo.

### 3.1.3 Factibilidad Económica

Al tener la mayor parte de la información vía web, nuestro gasto no es muy grande, ya que solo se usa dinero para copias y/o impresiones.

Tabla 14
*Factibilidad Económica*

| **Económicos** | |
|---|---|
| **Acceso a internet** | 320 |
| **Pasajes** | 600 |
| **Almuerzo** | 300 |
| **Copias** | 200 |
| **Libros** | 60 |

## 3.2. Modelado del negocio

Para iniciar con el desarrollo del modelo predictivo, antes debemos de entender cómo funciona el proceso del negocio que se desea automatizar, para tener la seguridad de que el modelo desarrollado cumpla con su finalidad, con el fin de lograr esto, se realizará un levantamiento de información detallado del negocio.

Target Organization Assessment - Estructura (organigrama)

**Descripción General de la Empresa**

Nuestra Universidad Autónoma del Perú se creó mediante Resolución N° 335-2007-CONAFU y con Resolución N° 171-2014-CONAFU acredita cumplir todas las normas legales y metas de su contrato social de creación. Tenemos siete años de excelencia legal y académica y seguiremos mejor. Ejecutamos convenios con universidades internacionales como Sao Paulo, Brasil, Internacional de Florida, Estados Unidos, Autónoma de Bucaramanga, Colombia y Global Earth, Costa Rica, y también de Perú como la Nacional de Trujillo, César Vallejo y Señor de Sipán, conformando con estas dos últimas el Consorcio Académico Universitario más grande del Perú.

**Nuestra Misión**

Formamos integralmente personas como agentes de cambio comprometidas con el desarrollo sostenible a través de la investigación, propuestas educativas innovadoras y altos estándares de calidad.

**Nuestra Visión**

Ser reconocida por su alta calidad académica, comprometida con la investigación, el desarrollo sostenible y acreditado internacionalmente

**Organigrama de la Empresa**

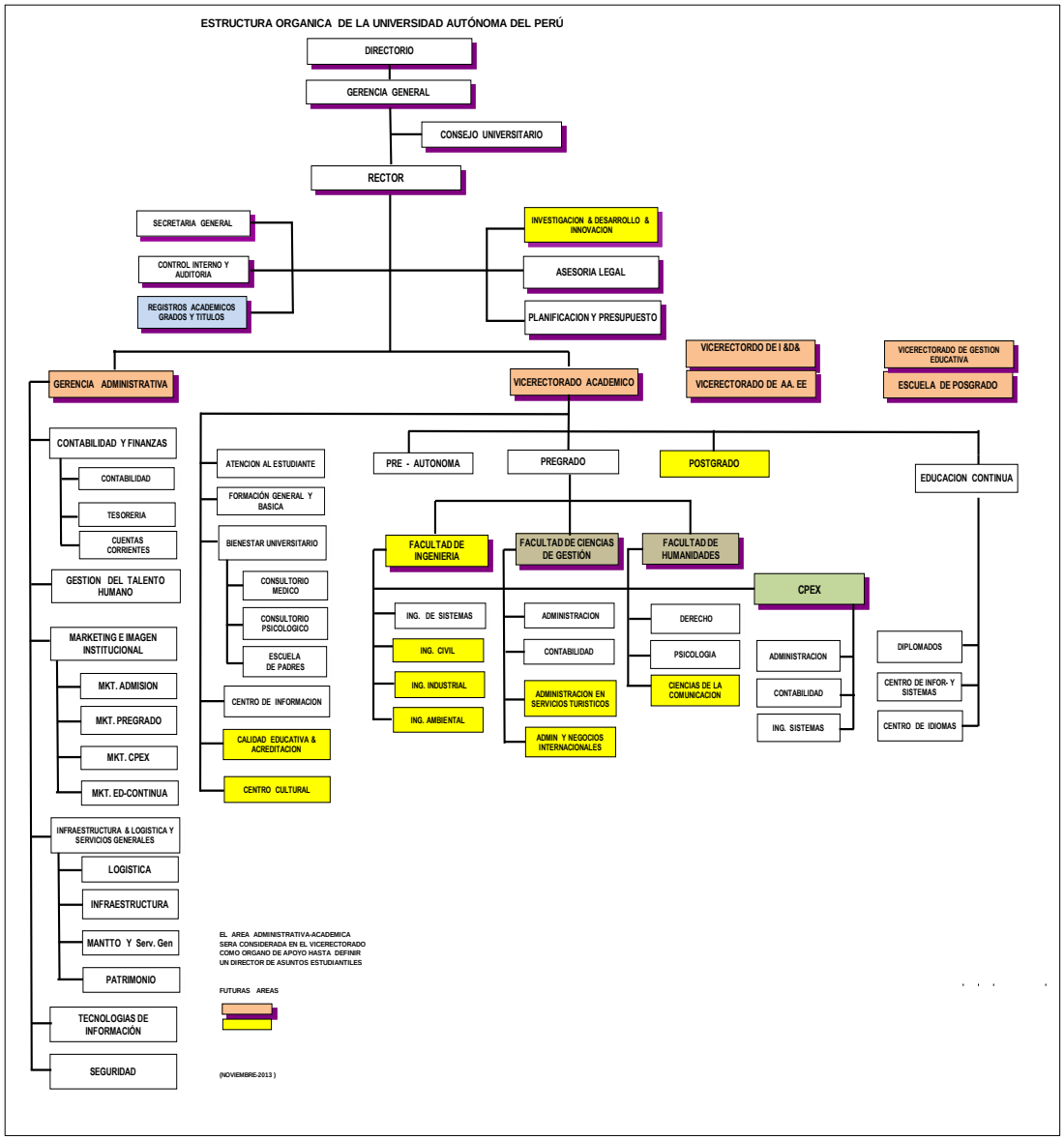


*Figura 17.* Diagrama de stakeholders (2016) adaptado de "Diagrama de Stakeholders" por Universidad Autónoma del Perú, 2016

## 3.2.1 Productos

La universidad carreras priorizando el entendimiento y aplicación de materias respetando las políticas establecidas autónoma del Perú se diferencia por su método de enseñanza en sus diferentes.

Tabla 15
*Productos y/o Servicios*

| **Carrera Profesional** | **Cursos Extracurriculares** | **CPEX** |
|---|---|---|
| Administración de Empresas | CISTEC | Administración de Empresas |
| Administración en Turismo y Hotelería | Centro de Idiomas | Derecho |
| Administración y Marketing | Cursos para la Gestión Publica | Psicología |
| Contabilidad | Cursos de Ofimática | Ingeniería de Sistemas |
| Economía y Finanzas | Redes y Telecomunicaciones | |
| Negocios Internacionales | Lenguaje de Programación | |
| Ciencias de la Comunicación | | |
| Derecho | | |
| Psicología | | |
| Arquitectura | | |
| Ingeniería Ambiental | | |
| Ingeniería Civil | | |
| Ingeniería de Sistemas | | |
| Ingeniería Industrial | | |
| Enfermería | | |
| Medicina Humana | | |
| Odontología | | |
| Terapia Física y Rehabilitación | | |

Tabla 16
*Productos y/o Servicios*

| **Postgrado** | **Pre** |
|---|---|
| Maestrías | |
| Programas de especialización | |

### 3.2.2 Stakeholders internos y externos

Tabla 17
*Stakeholders internos y externos*

| Matriz de Stakeholder | | | | |
|---|---|---|---|---|
| **TIPO** | **Expectativa** | **Cargo del Rol** | **Nivel de Interés** | **Nivel de Influencia** |
| **Interno** | Cumplimiento del crecimiento Universitario | Trabajadores del área administrativa | Alto | Alto |
| **Interno** | | Trabajadores del área de finanza | Alto | Alto |
| **Interno** | | Trabajadores del área de operaciones | Alto | Alto |
| **Interno** | | Trabajadores del área de RRHH | Alto | Alto |
| **Interno** | | Propietarios | Alto | Alto |
| **Interno** | | Directivos | Alto | Alto |
| **Externo** | | Ministerio de educación | Alto | Bajo |
| **Externo** | | Clientes | Alto | Bajo |
| **Externo** | | Consejo de rectores | Alto | Bajo |
| **Externo** | | Empresas | Alto | Bajo |
| **Externo** | | Centros de investigación | Alto | Bajo |
| **Externo** | | Gremios profesionales | Alto | Bajo |
| **Externo** | | Otras universidades | Alto | Bajo |
| **Externo** | | Partidos políticos | Alto | Bajo |
| **Externo** | | Fundaciones | Alto | Bajo |
| **Externo** | | Escuelas de Ed. Básica y secundaria | Alto | Bajo |
| **Externo** | | Municipios | Alto | Bajo |
| **Externo** | | Gobierno regional | Alto | Bajo |
| **Externo** | | Medios de comunicación | Alto | Bajo |
| **Externo** | | Ong | Alto | Bajo |
| **Externo** | | Asociaciones | Alto | Bajo |
| **Externo** | | Medio Ambiente. | Alto | Bajo |

Tabla 18
*Leyenda de Stakeholders*

| LEYENDA | | |
|---|---|---|
| **Tipo** | **Interés** | **Influencia** |
| Interno: Persona que se encuentra relacionada directamente con el desarrollo de la universidad.<br>Externo: Persona que no participa del desarrollo de la universidad, pero si puede llegar a hacer uso de él. | Alto: Personas que muestran un gran interés y están al pendiente del desarrollo de la universidad.<br>Bajo: Personas cuyo interés por el desarrollo de la universidad es mínimo. | Alto: Todas aquellas personas que pueden tener una influencia alta ya sea por su decisión, o el cargo que tengan, y puede ser tanto para beneficio o perjuicio del desarrollo de la universidad<br>Bajo: Personas cuyas decisiones no influyen positivamente o negativamente en el desarrollo de la universidad. |

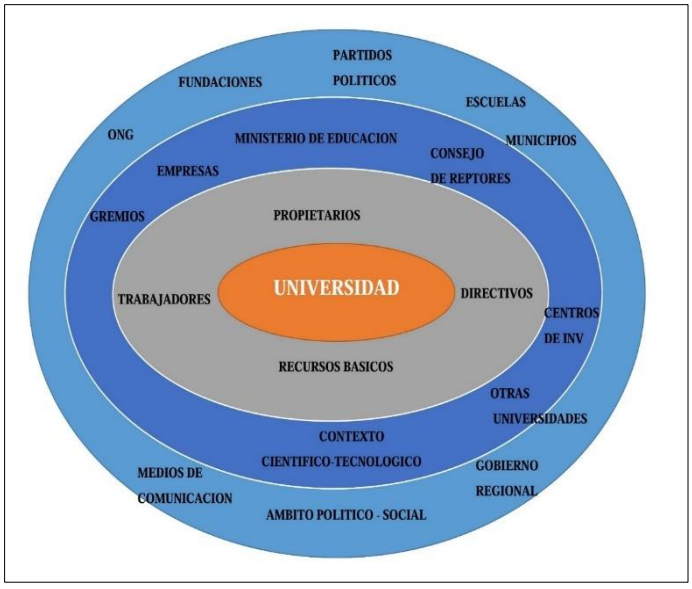


*Figura 18*. Cadena de valor

**Cadena de valor**

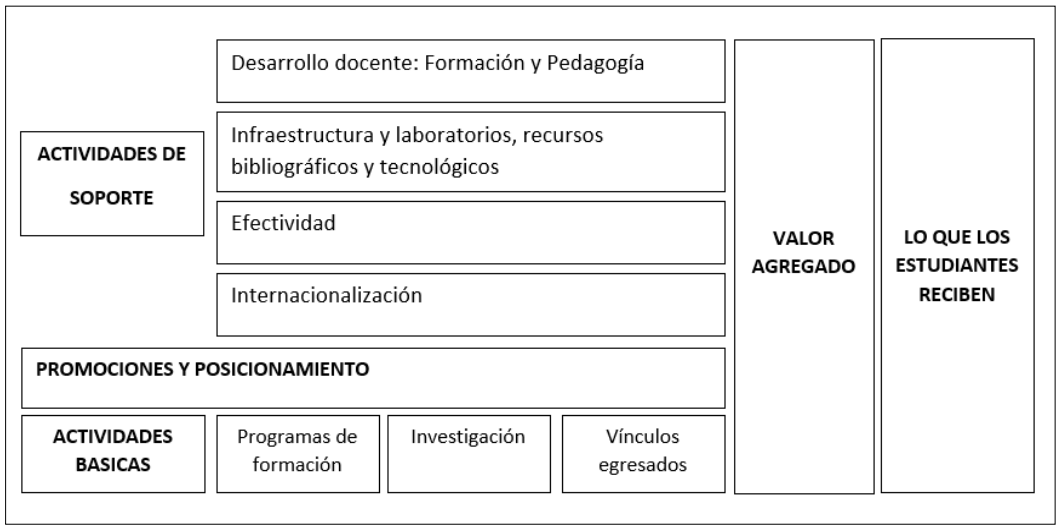


*Figura 19*. Cadena de valor

### 3.2.3 Identificación de procesos en CV

Proceso de actividades de soporte:

- Desarrollo: se busca el mejor trato alumno docente.
- Infraestructura: dar el mejor servicio brindando moderna infraestructura.
- Efectividad: garantizar el aprendizaje brindando cursos por dia de esta manera fomentamos la práctica de dicho tema.
- Internacionalización: fomentamos las relaciones profesionales y el crecimiento como persona, haciendo intercambios y congresos internacionales, de esta manera mejoramos el nivel del alumnado.

**Proceso de actividades básicas:**

- Formación: buscamos la mejor curricular para dar una óptima formación según la carrera deseada.
- Investigación: damos énfasis en la investigación, para que de esta manera se pueda generar nuevo conocimiento.
- Relación con el entorno: formamos profesionales con la capacidad de desarrollo en cualquier entorno, de esta manera el desempeño es óptimo.
- Hacen reuniones con egresados para poder dar a conocer su experiencia cuando eran estudiantes y ahora como profesionales.

### 3.2.4 Procesos de negocios


*Figura 20.* Proceso de Negocio

## 3.3. Modelado del Proceso

### 3.3.1 Modelado de contexto

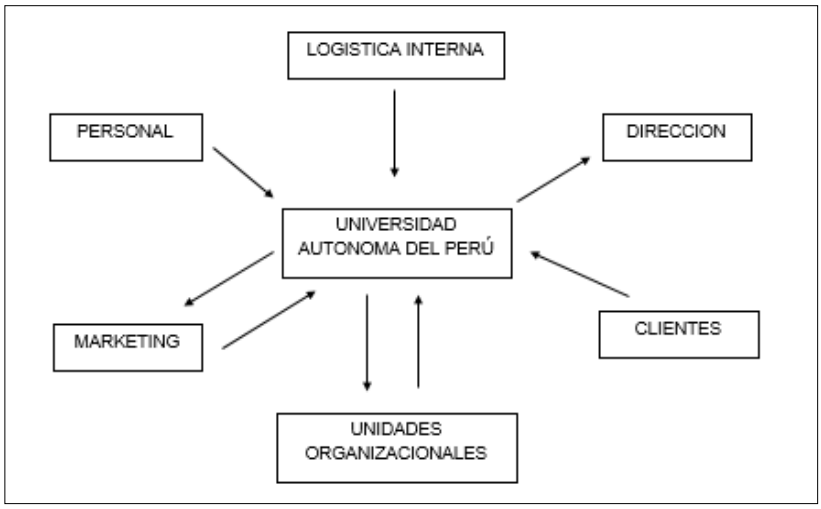


*Figura 21*. Modelo de contexto

## 3.4 Incepción del Proyecto

### 3.4.1 Presentación

En esta parte de la investigación pasamos a la parte más práctica, donde iremos aplicando cada una de las fases de la metodología CRISP-DM al problema práctico que nos planteamos, que es la extracción y explotación de datos del entorno universitario del área financiera.

Iremos numerando cada una de las fases de la metodología tal y como están numeradas en el documento original

### 3.4.2 Comprensión del Negocio

A continuación, iremos siguiendo cada una de las tareas de las que consta esta primera fase en el proceso de la minería de datos, cuya finalidad es determinar los objetivos y requisitos de la investigación desde una perspectiva de negocio, para más adelante poder convertirlos en objetivos desde el punto de vista técnico.

#### 3.4.2.1 Determinar los Objetivos del Negocio

El objetivo de la minería de datos que se va a aplicar en esta investigación es el de hacer predicciones lo más fiables posible a partir de los datos de los que ya se disponen de los alumnos en una universidad. El objetivo es proporcionar un mejor servicio de enseñanza a los alumnos y así poder captar más alumnos para que realicen sus estudios en la universidad.

#### 3.4.2.2 Contexto

En referencia a la situación de negocio en la organización (universidad) al principio de esta investigación se puede decir que se cuenta con una base de datos de los alumnos actualmente cursando las carreras que se encuentran definidas en ella. Sin embargo, no existe ningún estudio en profundidad sobre el comportamiento de los estudiantes de los que se puedan sacar conclusiones o patrones para hacer predicciones sobre los futuros estudiantes que puedan ser morosos.

#### 3.4.2.3 Objetivos del negocio

Los objetivos del negocio como ya se ha mencionado son la predicción de datos para los alumnos de nuevo ingreso como también los de curricula actual de tal manera que se pueda hacer una estimación fiable partiendo de los datos que ya tenemos de dichos alumnos. Se podrían hacer muchas predicciones según las necesidades de la universidad en cada momento, pero en esta investigación se han definido los siguientes objetivos:

- Determinar un modelo predictivo para clasificar a los posibles alumnos morosos
- Realizar un estudio cualitativo del comportamiento de las personas.
- Realizar una investigación del comportamiento de las personas para generar conocimiento sobre los riesgos en los otorgamientos de facilidades.

- Determinar los beneficios que se obtendrán al aplicar el modelo predictivo.
- Reconocer los beneficios de usar Data Mining para la obtención del modelo predictivo.

Estos informes pueden ser muy útiles para los del área de finanzas a la hora de tomar la decisión si el alumno puede ser acto para un descuento de deuda o un refinanciamiento, así como para detectar aquellas personas morosas y de esta forma intentar averiguar por qué muchos alumnos dejan la universidad, ya sea por falta de dinero para cancelar todas sus deudas, deuda acumulada, etc. Todo esto permitirá a la universidad mejorar la calidad de los servicios ofrecidos a los estudiantes.

#### 3.4.2.4 Criterios de éxito del negocio

Desde el punto de vista del negocio se establece como criterio de éxito la posibilidad de realizar predicciones sobre nuevos alumnos con un elevado porcentaje de fiabilidad, de tal forma que se puedan dar un mejor resultado al momento de querer saber si un alumno es moroso. Otro criterio de éxito del negocio sería disminuir el porcentaje de alumnos morosos, ya que esto conllevaría a una mejora de trabajo.

### 3.4.3 Evaluación de la Situación

Se cuenta con una base de datos mysql con información detallada de los alumnos han sugerido un refinanciamiento o descuento de deuda, por lo que se puede afirmar que se dispone de una cantidad de datos más que suficiente para poder resolver el problema. Esta información incluye el acceso mediante el sistema que manejan al registrar los casos de los estudiantes solicitando descuento.

#### 3.4.3.1 Inventario de recursos

En cuanto a recursos de software disponemos del programa de minería de datos Weka que proporciona herramientas para realizar tareas de minería de datos sobre la base de datos con la que contamos para el almacenamiento de los datos.

Los recursos de hardware de los que disponemos son un ordenador de sobremesa con las siguientes características:

- Marca: HP
- Modelo: HP14
- Memoria RAM: 4 GB
- Capacidad de almacenamiento: 750 GB

- Tarjeta gráfica: AMD RADEON R5 1 GB
- Sistema operativo: WINDOWS 10
- Monitor TFT: VGA

La fuente de datos es una base de datos mysql con la información de los alumnos de las cuales solicitaron un refinanciamiento o un descuento por parte de la Universidad y a su vez nos muestra el historial de todos sus pagos hasta el momento.

#### 3.4.3.2 Requisitos, supuestos y restricciones

Al no poder utilizar los datos personales de alumnos reales debido a cuestiones legales, se ha tenido que utilizar una base de datos con datos recolectados de alumnos de la Universidad Autónoma del Perú.

#### 3.4.3.3 Terminología

Ver Anexo 1: Glosario de terminología de minería de datos.

#### 3.4.3.4 Costes y beneficios

Los datos de este proyecto no suponen ningún coste adicional a la universidad ya que estos datos perteneces a la propia universidad desde el momento en el que el alumno se matricula en ella.

En cuanto a beneficios, se puede decir que este proyecto si genere algún beneficio económico para la universidad directamente y a su vez también indirectamente ya que el objetivo de este proyecto es mejorar la calidad de los servicios ofrecidos a los alumnos por parte de la universidad, y por tanto la satisfacción de los clientes (los alumnos), y esto se traduce en prestigio para la universidad.

### 3.4.4 Determinar los Objetivos de la Minería de Datos

- Los objetivos en términos de minería de datos son:
- Determinar un modelo predictivo para clasificar a los posibles alumnos morosos
- Realizar un estudio cualitativo del comportamiento de las personas.
- Realizar una investigación del comportamiento de las personas para generar conocimiento sobre los riesgos en los otorgamientos de facilidades.
- Determinar los beneficios que se obtendrán al aplicar el modelo predictivo.
- Reconocer los beneficios de usar Data Mining para la obtención del modelo predictivo.

#### 3.4.4.1 Criterios de éxito de minería de datos

Desde el punto de vista de la minería de datos se establece como criterio de éxito la posibilidad de realizar predicciones sobre nuevos alumnos con un elevado porcentaje de fiabilidad, concretamente definimos este porcentaje en un 100%. El grado de fiabilidad lo determinará el algoritmo específico que se emplee a la hora de conseguir el modelo de la minería de datos, por lo que este tema se volverá a abordar más adelante en el paso 5 de la metodología (evaluación).

#### 3.4.4.2 Realizar el Plan del Proyecto

El proyecto se dividirá en las siguientes etapas para facilitar su organización y estimar el tiempo de realización del mismo:

**Etapa 1:** Análisis de la estructura de los datos y la información de la base de datos. Tiempo estimado: 2 semanas.

**Etapa 2:** Ejecución de consultas para tener muestras representativas de los datos. Tiempo estimado: 1 semana.

**Etapa 3:** Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la minería de datos sobre ellos. Tiempo estimado: 3 semanas.

**Etapa 4:** Elección de las técnicas de modelado y ejecución de las mismas sobre los datos. Tiempo estimado: 1 semana.

**Etapa 5:** Análisis de los resultados obtenidos en la etapa anterior, si fuera necesario repetir la etapa 4. Tiempo estimado: 1 semana.

**Etapa 6:** Producción de informes con los resultados obtenidos en función de los objetivos de negocio y los criterios de éxito establecidos. Tiempo estimado: 1 semana.

**Etapa 7:** Presentación de los resultados finales. Tiempo estimado: 1 semana.

Nota: en paralelo a la realización de cada una de estas etapas se irá construyendo el diccionario de terminología de minería de datos (Anexo 1).

#### 3.4.4.3 Evaluación inicial de herramientas y técnicas

La herramienta que se va a utilizar para llevar a cabo este proyecto de minería de datos es el weka ya que esta herramienta se adapta bien a la metodología que estamos

empleando y sobre todo a la base de datos en la que están almacenados todos los datos de los estudiantes. Además, gracias a esta herramienta no necesitamos pasar la información almacenada en la base de datos.

En cuanto a las técnicas que se van a emplear para la extracción de conocimiento, weka nos ofrece los siguientes tipos de tareas de minería de datos:

• Predictivas o Clasificación o Regresión • Descriptivas o Agrupamiento (clustering) o Reglas de asociación

Weka además utiliza los siguientes algoritmos para resolver los problemas: árboles de decisión, clasificador bayesiano naive, SVM (Máquinas de Vectores de Soporte) y GLM (Modelo Lineal Generalizado).

Los árboles de decisión, también llamados modelos basados en árboles se fundamentan en el principio de "divide y vencerás". Los árboles se van construyendo con nodos, de tal forma que en cada nodo se establecen unas condiciones sobre uno o varios atributos, dividiendo de esta manera el conjunto de casos en subconjuntos que cumplen las condiciones. Estos subconjuntos a su vez se vuelven a dividir añadiendo más niveles al árbol hasta detenerse en el punto en el que se cumpla algún criterio.

El clasificador bayesiano naive es un clasificador probabilístico basado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el nombre de naive (ingenuo).

Las Máquinas de Vectores de Soporte o Support Vector Machines (SVM) son un conjunto de algoritmos de aprendizaje supervisado que están principalmente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Las nuevas muestras se clasifican en una u otra clase en función de la proximidad con el modelo producido. Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en

problemas de clasificación o regresión. Una buena separación entre las clases permite una clasificación mejor.

El Modelo Lineal Generalizado o Generalized Linear Model (GLM) es una generalización de la regresión de mínimos cuadrados ordinaria. Relaciona la distribución aleatoria de la variable dependiente en el experimento (la función de distribución) con la parte sistemática (no aleatoria) a través de una función llamada la función de enlace. Los modelos lineales generalizados fueron formulados como una manera de unificar varios modelos estadísticos, incluyendo la regresión lineal, regresión logística y regresión de Poisson, bajo un solo marco teórico. Esto ha permitido desarrollar un algoritmo general para la estimación de máxima verosimilitud en todos estos modelos

#### 3.4.4.4 Comprensión de los Datos

En esta segunda fase de la metodología CRISP-DM se realiza la recolección inicial de los datos para poder establecer un primer contacto con el problema, familiarizarse con los datos y averiguar su calidad, así como identificar las relaciones más evidentes para formular las primeras hipótesis.

#### 3.4.4.5 Recolectar los Datos Iniciales

Los datos utilizados en esta investigación son datos referentes a alumnos que incluyen información personal sobre ellos como puede ser, su edad, ciclo, si trabajan, etc., por lo que no hemos podido utilizar datos reales de acuerdo a las encuestas realizadas hacia los alumnos. Debido a la gran cantidad de registros que son necesarios para poder hacer un trabajo de minería de datos con éxito, la opción de insertar estos registros manualmente uno a uno en la base de datos no era viable, por lo que se optó por crear un programa en el lenguaje de programación Java, cuya salida fueran los distintos scripts de inserción de datos (uno por cada tabla).

**A continuación, listamos los datos adquiridos:**

**Deudas**

Cada deuda está identificada por un número. Toda deuda está relacionada con un pago cual pertenece.

**Pagos**

Cada pago que la universidad asigna a los alumnos de acuerdo a su categoría está identificado por un número.

**Refinanciamientos**

Cada refinanciamiento está también identificado por un número.

**Alumnos**

Cada alumno está identificado por su id de alumno que es un valor numérico. Todo alumno está relacionado con un pago y con una deuda que es la que el alumno pagara en la universidad.

**Fechas**

Las fechas son extraídas en formato numérico con el formato caaaa, donde c es el número del cuatrimestre en cuestión (1 para el primer cuatrimestre, 2 para el segundo, y 3 para la convocatoria extraordinaria), y aaaa son los cuatro dígitos del año al que se refiere. Así, 21998 se referiría al segundo cuatrimestre del año 1998.

Los atributos específicos que serán útiles a la hora de hacer la minería de datos son:

- Identificador de alumno
- Centro de procedencia del alumno
- Identificador del Pago
- Historial de los pagos
- Tiempo (ciclos)
- Identificador de la deuda
- Fecha en la que se generó la deuda
- Cumplimiento de pago de refinanciamiento

Las tablas en las que se recogen los datos necesarios para la minería de datos son:

- Tiempo
- Riesgo
- Alumno
- Deuda
- Ciclo

### 3.4.4.6 Descripción de los datos

Los datos fueron creados por los autores de esta tesis, ya que por motivos normativos la universidad no nos pudo facilitar aun la base de datos, dicho esto se presenta el cuadro de los datos a utilizar.

Tabla 19
*Descripción de Datos*

| TIPO | ATRIBUTOS | POSIBLES VALORES |
|---|---|---|
| **INDIVIDUALES** | Solvencia | Apoyo Economico |
| **INSTITUCIONAL** | Tarjeta | Si,no |
| **SOCIECONOMICO** | Tipo de ahorro | Mensual, semanal, Diario |
| **SOCIECONOMICO** | Prestamos | Todos los posibles |
| **OTRO** | Asistenta Social | Si, no |
| **INDIVIDUALES** | Vivienda | todos los posibles |
| **OTRO** | Hermanos | Todos los posibles |
| **SOCIECONOMICO** | Ingreso Mensual | Categorias |
| **OTRO** | Salidas | Si, no |
| **SOCIECONOMICO** | Mensualidad | Categorias |

### Registros generados

Aparte de estas dos operaciones, no ha sido necesario generar nuevos atributos ni integrar nuevos registros a la base de datos ya que ésta está completa y ha sido creada específicamente para su uso en esta investigación

### 3.4.5 Integrar los Datos

No ha sido necesaria la creación de nuevas estructuras (campos, registros, etc.), ni la fusión entre distintas tablas de la base de datos, ya que el programa Weka se encarga de realizar estas tareas automáticamente sin que el usuario tenga que crear nuevas tablas, registros o campos manualmente.

### 3.4.6 Formatear los Datos

El campo con la información referente al nivel de morosidad de los alumnos ha sido codificado con valores numéricos ya que la herramienta de minería de datos exige que los datos a estudiar sean numéricos, para dar un mejor trabajo a la hora de hacer la minería se colocó en escala de intervalos del 1 al 100% dando una pequeña referencia de los valores y su significado.

Se propuso de la siguiente forma:

- 10% → Baja probabilidad de ser moroso
- 20% → Baja probabilidad de ser moroso
- 30% → Baja a media probabilidad de ser moroso
- 40% → Baja a media probabilidad de ser moroso
- 50% → Media probabilidad de ser moroso
- 60% → Media probabilidad de ser moroso
- 70%→ Alta probabilidad de ser moroso
- 80% → Alta probabilidad de ser moroso
- 90% → Alta probabilidad de ser moroso
- 100% → Alta probabilidad de ser moroso

Se hace el cambio al momento de usar los datos para evitar conflictos de la siguiente manera:

No es necesario cambiar el orden de ningún campo dentro de los registros, ni tampoco la reordenación de los registros dentro de las tablas. Tampoco es necesario cambiar el formato de ninguno de los campos que se van a utilizar para la minería de datos ya que el formato actual es admitido por la herramienta Weka.

**Modelado**

En esta fase de la metodología se escogerá la técnica (o técnicas) más apropiadas para los objetivos marcados de la minería de datos. A continuación, y una vez realizado un plan de prueba para los modelos escogidos, se procederá a aplicar dichas

técnicas sobre los datos para generar el modelo y por último se tendrá que evaluar si dicho modelo ha cumplido los criterios de éxito o no.

### 3.5 Escoger la Técnica de Modelado

Debido a que se va a utilizar el software Weka para realizar la minería de datos, deberemos utilizar alguna de las técnicas de modelado que nos ofrece esta herramienta de acuerdo con los objetivos de nuestra investigación.

De los modelos que nos ofrece Weka, el que mejor se adapta a nuestros objetivos sería un modelo de clasificación, puesto que los problemas que queremos resolver son problemas de predicción y los campos que se quieren predecir contienen valores continuos, acoplándose de mejor manera una de las técnicas del modelo ya dicho, que es la técnica del árbol.

#### 3.5.1 Generar el Plan de Prueba

El procedimiento que se empleará para probar la calidad y validez del modelo será el de utilizar las métricas de la matriz de confusión (matrix confusion), y la validación cruzada (k-CV: k-fold Cross-Validation). Para entender mejor estos indicadores vamos a describirlos a continuación.

Matriz de confusión: Mediante esta métrica de validación se logra obtener la precisión del clasificador

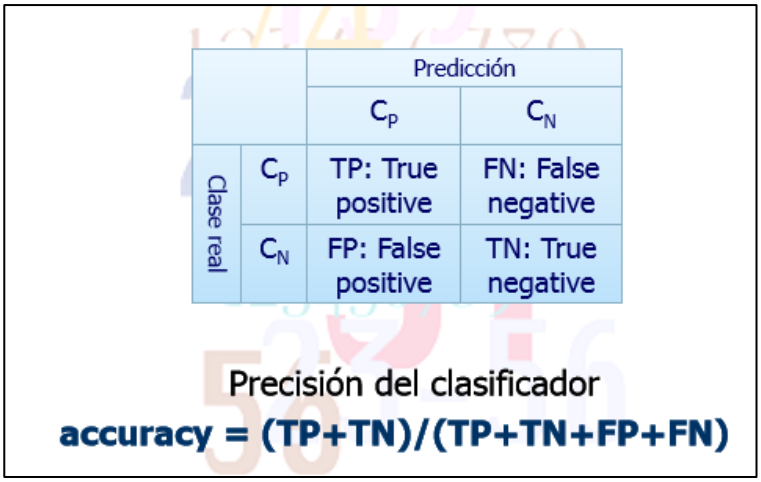


*Figura 22*. Matriz de confusión

Validación Cruzada, mediante este método se divide aleatoriamente el conjunto de datos en k subconjuntos de intersección vacía (más o menos del mismo tamaño).

### 3.5.2 Construir el Modelo

A continuación, se procederá a ejecutar el modelo elegido sobre los datos de entrenamiento. En este apartado se describirán los ajustes de parámetros del modelo que se eligen en la herramienta de minería de datos, así como la salida de dicho modelo y su descripción.

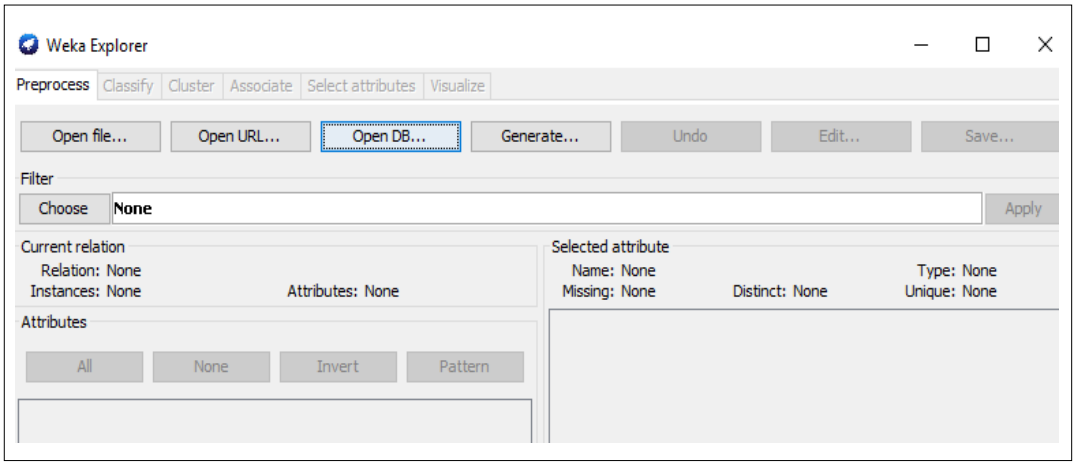


*Figura 23.* Entorno weka

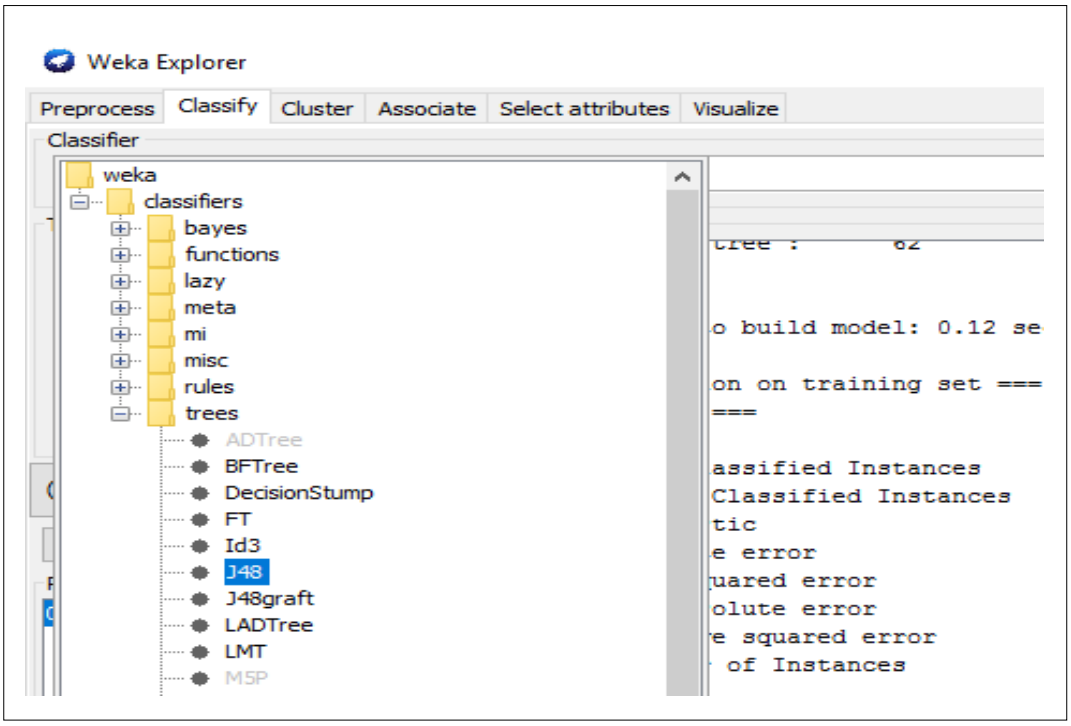


*Figura 24.* Entorno Clasificación

## Modelo 1

### Atributo solvencia

Classifier output

```
J48 pruned tree
------------------

ayuda_social = si
|   frec_ahorro = mensual
|   |   sexo = hombre
|   |   |   vivienda = propia
|   |   |   |   prestamo = aval
|   |   |   |   |   ingreso_mensual = a: otros (57.0/33.0)
|   |   |   |   |   ingreso_mensual = b: padres (24.0/12.0)
|   |   |   |   |   ingreso_mensual = c: padres (38.0/20.0)
|   |   |   |   prestamo = empeño: otros (0.0)
|   |   |   |   prestamo = prestamo: otros (13.0)
|   |   |   vivienda = alquilado: solo (23.0/12.0)
|   |   |   vivienda = otros: solo (54.0/24.0)
|   |   sexo = mujer
|   |   |   prestamo = aval: padres (37.0/7.0)
|   |   |   prestamo = empeño: padres (0.0)
|   |   |   prestamo = prestamo
|   |   |   |   vivienda = propia: solo (26.0/13.0)
|   |   |   |   vivienda = alquilado
|   |   |   |   |   mensualidad = a
|   |   |   |   |   |   ingreso_mensual = a: solo (8.0/4.0)
|   |   |   |   |   |   ingreso_mensual = b: otros (10.0/4.0)
|   |   |   |   |   |   ingreso_mensual = c: otros (7.0/2.0)
|   |   |   |   |   mensualidad = b
|   |   |   |   |   |   hermanos = 1hermano: otros (9.0/3.0)
|   |   |   |   |   |   hermanos = 2hermano: solo (5.0/3.0)
|   |   |   |   |   |   hermanos = 3hermano: otros (0.0)
|   |   |   |   |   |   hermanos = masde3hermano: otros (0.0)
|   |   |   |   |   mensualidad = c: solo (8.0/4.0)
|   |   |   |   |   mensualidad = d: otros (21.0/12.0)
```

*Figura 25*. Entorno Data Modelo 2

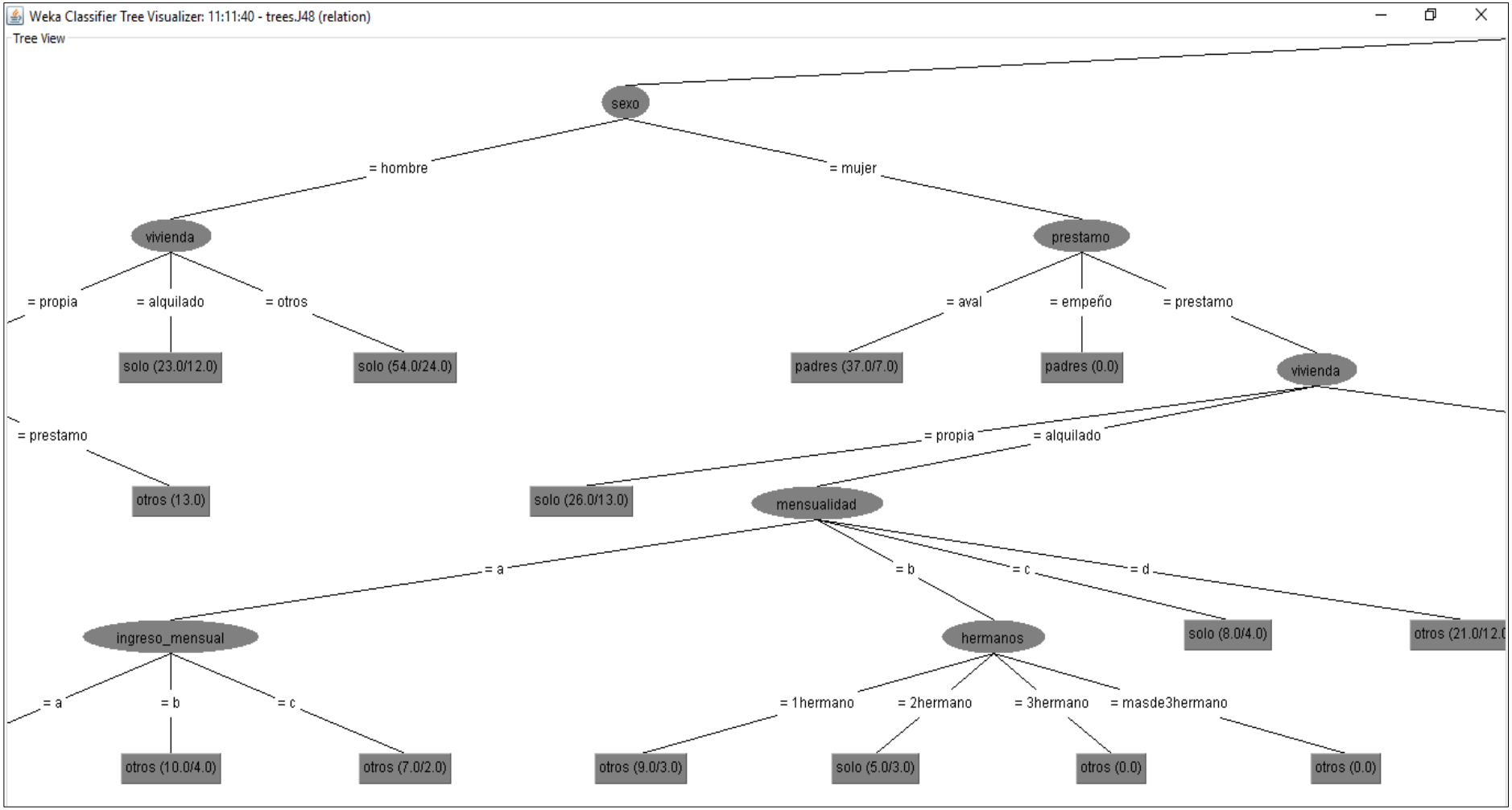


*Figura 26*. Entorno Árbol de decisión 1

## Modelo 2

## Atributo frec_ahorro

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

sexo = hombre
|   prestamo = aval
|   |   vivienda = propia
|   |   |   ingreso_mensual = a: mensual (100.0)
|   |   |   ingreso_mensual = b: mensual (32.0)
|   |   |   ingreso_mensual = c
|   |   |   |   mensualidad = a
|   |   |   |   |   ayuda_social = si: semanal (2.0)
|   |   |   |   |   ayuda_social = no: mensual (30.0)
|   |   |   |   mensualidad = b: mensual (4.0)
|   |   |   |   mensualidad = c: mensual (37.0/3.0)
|   |   |   |   mensualidad = d: semanal (10.0)
|   |   vivienda = alquilado
|   |   |   mensualidad = a: semanal (14.0)
|   |   |   mensualidad = b: semanal (0.0)
|   |   |   mensualidad = c: mensual (1.0)
|   |   |   mensualidad = d: mensual (3.0)
|   |   vivienda = otros: mensual (31.0/2.0)
|   prestamo = empeño
|   |   ayuda_social = si: diario (8.0)
|   |   ayuda_social = no: mensual (138.0)
|   prestamo = prestamo
|   |   ingreso_mensual = a
|   |   |   hermanos = 1hermano
|   |   |   |   vivienda = propia: mensual (4.0)
|   |   |   |   vivienda = alquilado: mensual (26.0)
|   |   |   |   vivienda = otros: semanal (18.0/3.0)
```

*Figura 27*. Entorno Data Modelo 2

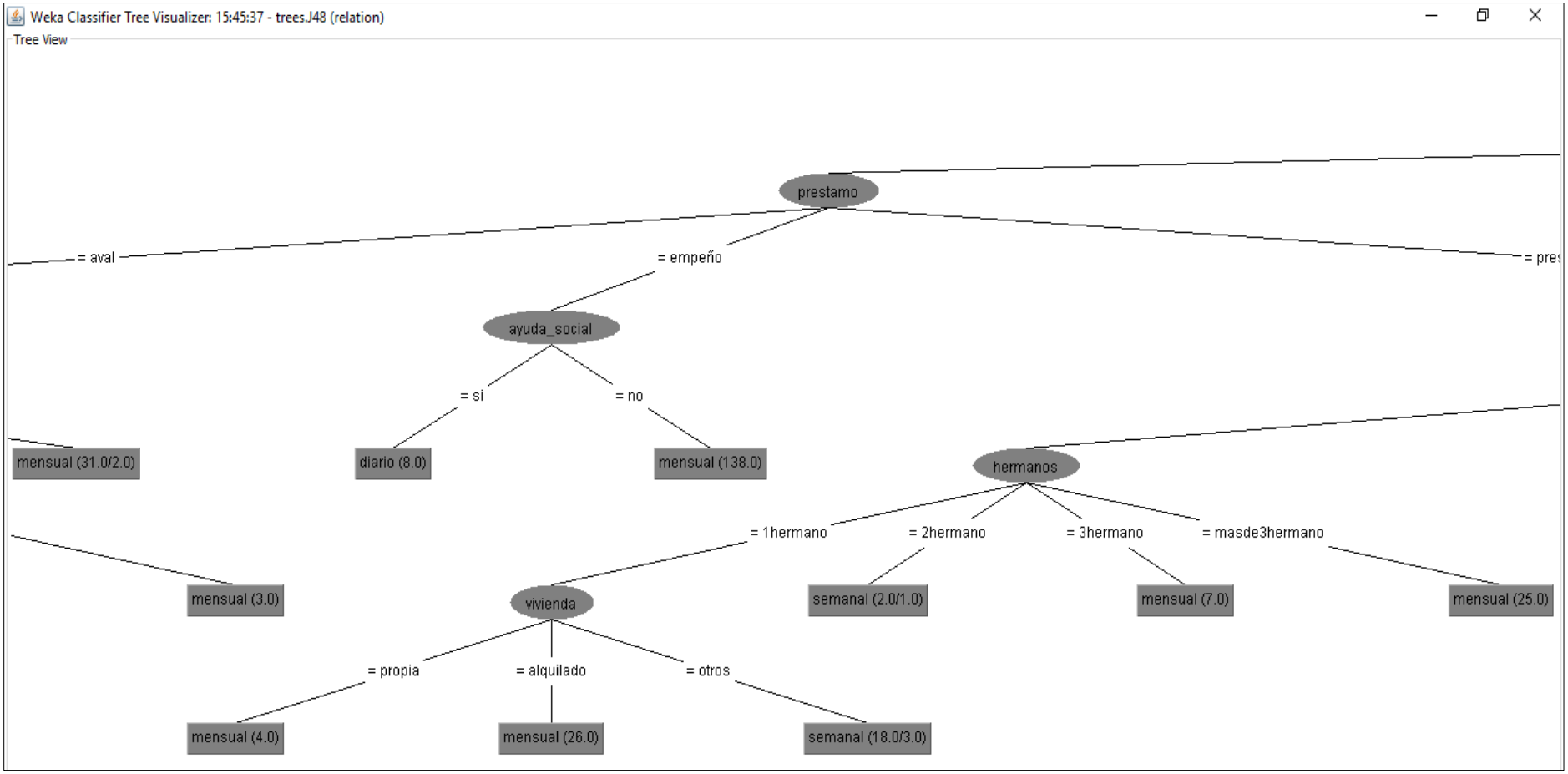


*Figura 28.* Entorno Árbol de decisión Modelo 2

## Modelo 3

## Atributo sexo

```
J48 pruned tree
------------------

ayuda_social = si
|   frec_ahorro = mensual
|   |   prestamo = aval: hombre (185.0/37.0)
|   |   prestamo = empeño: hombre (0.0)
|   |   prestamo = prestamo
|   |   |   ingreso_mensual = a
|   |   |   |   hermanos = 1hermano
|   |   |   |   |   solvencia = solo
|   |   |   |   |   |   vivienda = propia: mujer (4.0)
|   |   |   |   |   |   vivienda = alquilado: hombre (8.0)
|   |   |   |   |   |   vivienda = otros: hombre (2.0)
|   |   |   |   |   solvencia = padres: mujer (27.0/8.0)
|   |   |   |   |   solvencia = otros
|   |   |   |   |   |   vivienda = propia: hombre (4.0)
|   |   |   |   |   |   vivienda = alquilado: mujer (14.0/3.0)
|   |   |   |   |   |   vivienda = otros: mujer (2.0)
|   |   |   |   hermanos = 2hermano: mujer (12.0)
|   |   |   |   hermanos = 3hermano: mujer (0.0)
|   |   |   |   hermanos = masde3hermano
|   |   |   |   |   vivienda = propia
|   |   |   |   |   |   solvencia = solo: mujer (8.0)
|   |   |   |   |   |   solvencia = padres: mujer (8.0)
|   |   |   |   |   |   solvencia = otros: hombre (3.0)
|   |   |   |   |   vivienda = alquilado: mujer (0.0)
|   |   |   |   |   vivienda = otros
|   |   |   |   |   |   solvencia = solo: hombre (14.0)
|   |   |   |   |   |   solvencia = padres: hombre (14.0/7.0)
|   |   |   |   |   |   solvencia = otros: mujer (15.0/1.0)
```

*Figura 29.* Entorno Data Modelo 3

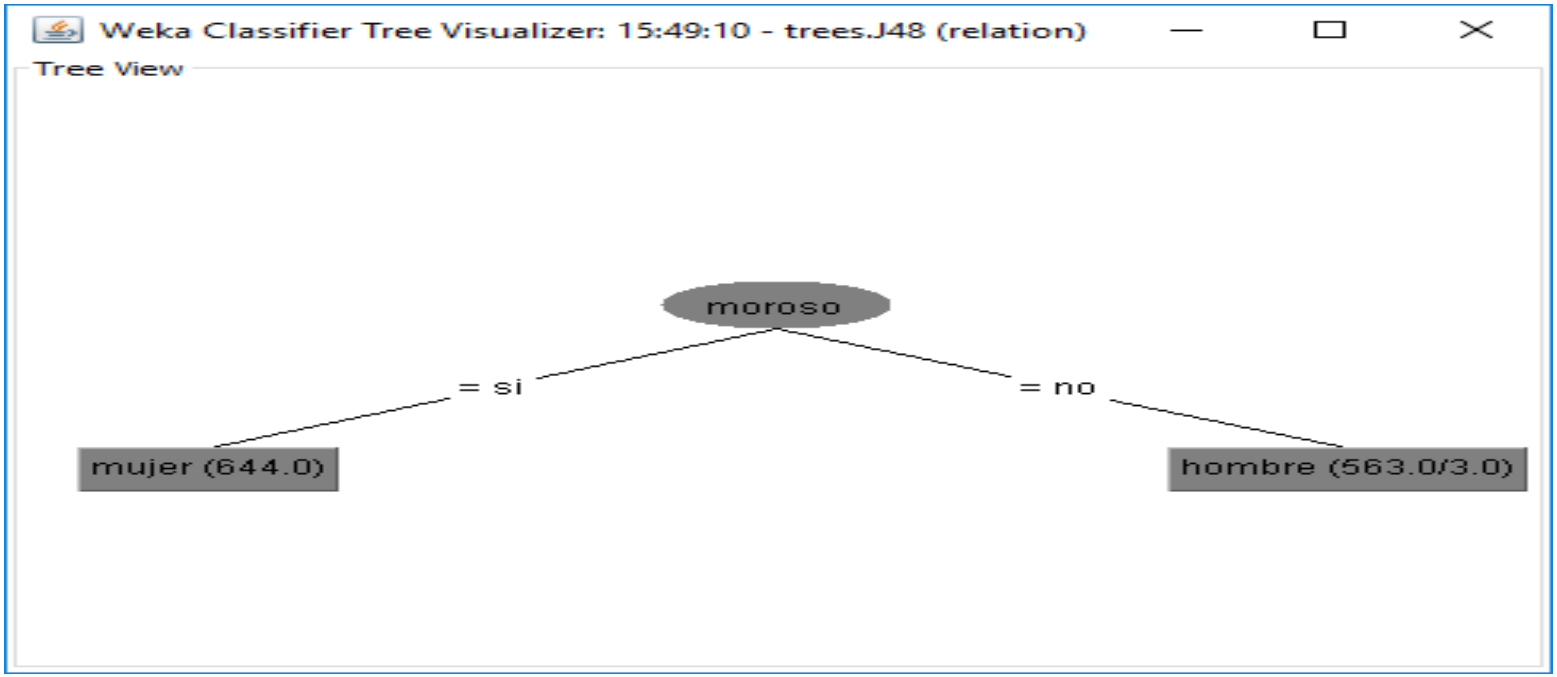


*Figura 30*. Entorno Árbol de decisión Modelo 3

## Modelo 4

## Atributo préstamo

```
------------------

ayuda_social = si
|   frec_ahorro = mensual
|   |   hermanos = 1hermano
|   |   |   vivienda = propia
|   |   |   |   mensualidad = a: prestamo (9.0)
|   |   |   |   mensualidad = b: aval (37.0)
|   |   |   |   mensualidad = c: aval (0.0)
|   |   |   |   mensualidad = d
|   |   |   |   |   salir_fin_semana = si: aval (24.0)
|   |   |   |   |   salir_fin_semana = no: prestamo (6.0)
|   |   |   vivienda = alquilado
|   |   |   |   salir_fin_semana = si
|   |   |   |   |   mensualidad = a: prestamo (6.0)
|   |   |   |   |   mensualidad = b: aval (0.0)
|   |   |   |   |   mensualidad = c: aval (0.0)
|   |   |   |   |   mensualidad = d: aval (6.0)
|   |   |   |   salir_fin_semana = no: prestamo (58.0/4.0)
|   |   |   vivienda = otros
|   |   |   |   ingreso_mensual = a: prestamo (15.0/2.0)
|   |   |   |   ingreso_mensual = b: aval (31.0)
|   |   |   |   ingreso_mensual = c: prestamo (4.0)
|   |   hermanos = 2hermano
|   |   |   sexo = hombre: aval (37.0/1.0)
|   |   |   sexo = mujer: prestamo (23.0)
|   |   hermanos = 3hermano: aval (45.0)
|   |   hermanos = masde3hermano: prestamo (75.0)
|   frec_ahorro = semanal
|   |   vivienda = propia
|   |   |   ingreso_mensual = a
|   |   |   |   hermanos = 1hermano: aval (12.0)
|   |   |   |   hermanos = 2hermano
```

*Figura 31*. Entorno Data Modelo 4

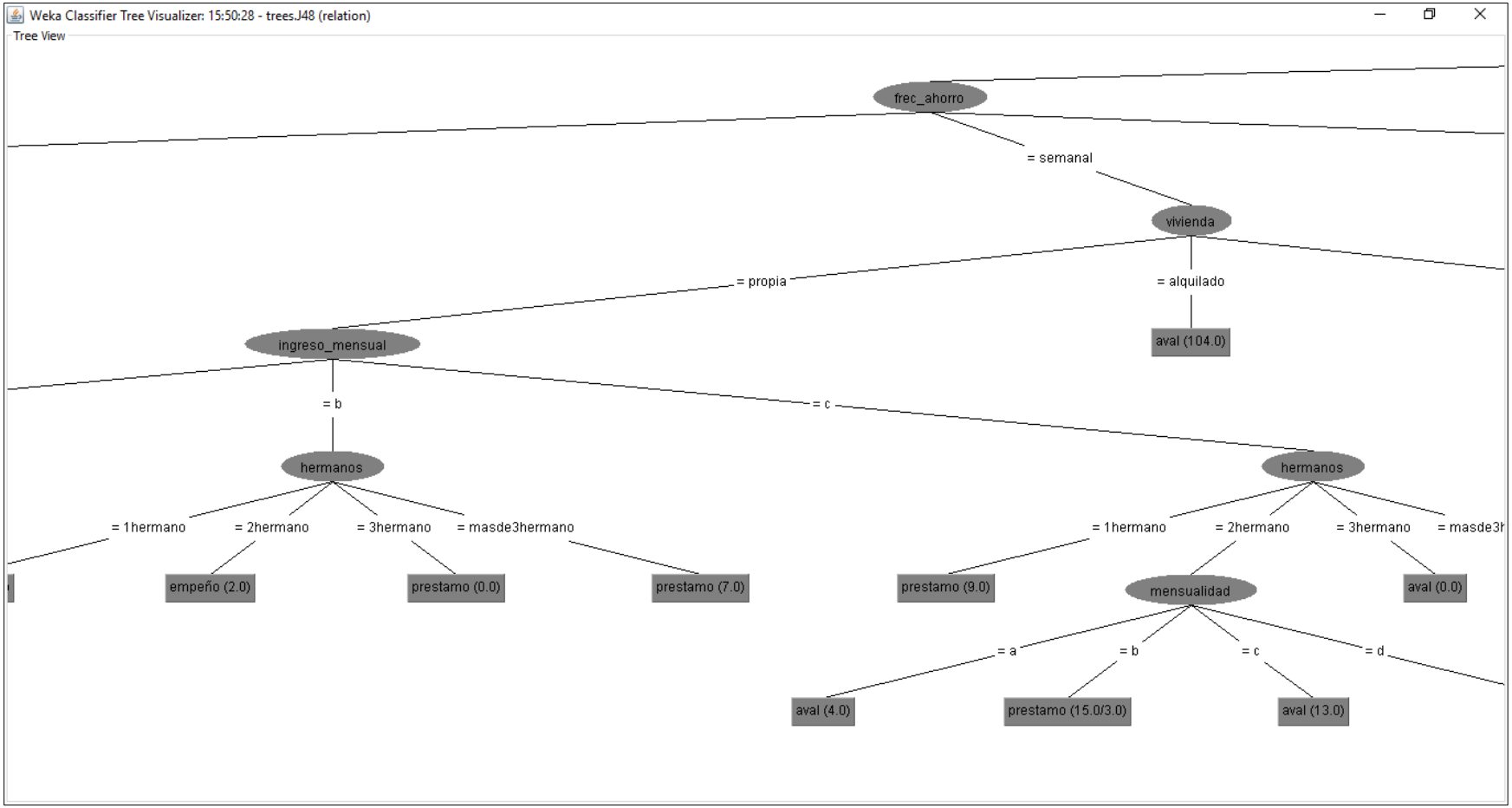


*Figura 32.* Entorno Árbol de decisión Modelo 4

## Modelo 5

## Atributo préstamo

Classifier output

```
sexo = hombre
|   prestamo = aval
|   |   mensualidad = a
|   |   |   vivienda = propia
|   |   |   |   frec_ahorro = mensual: no (46.0/4.0)
|   |   |   |   frec_ahorro = semanal: si (2.0)
|   |   |   |   frec_ahorro = diario: no (0.0)
|   |   |   vivienda = alquilado
|   |   |   |   salir_fin_semana = si: si (8.0)
|   |   |   |   salir_fin_semana = no
|   |   |   |   |   ingreso_mensual = a: si (0.0)
|   |   |   |   |   ingreso_mensual = b: no (2.0)
|   |   |   |   |   ingreso_mensual = c: si (4.0)
|   |   |   vivienda = otros: si (16.0)
|   |   mensualidad = b
|   |   |   hermanos = 1hermano: si (30.0)
|   |   |   hermanos = 2hermano: no (24.0)
|   |   |   hermanos = 3hermano: si (0.0)
|   |   |   hermanos = masde3hermano: si (0.0)
|   |   mensualidad = c
|   |   |   hermanos = 1hermano: no (8.0)
|   |   |   hermanos = 2hermano: si (39.0)
|   |   |   hermanos = 3hermano: si (29.0)
|   |   |   hermanos = masde3hermano: si (0.0)
|   |   mensualidad = d
|   |   |   hermanos = 1hermano: si (29.0)
|   |   |   hermanos = 2hermano
|   |   |   |   frec_ahorro = mensual: no (5.0)
|   |   |   |   frec_ahorro = semanal: si (10.0)
|   |   |   |   frec_ahorro = diario: si (0.0)
|   |   |   hermanos = 3hermano: si (12.0/6.0)
|   |   |   hermanos = masde3hermano: si (0.0)
```

*Figura 33*. Entorno Data Modelo 5

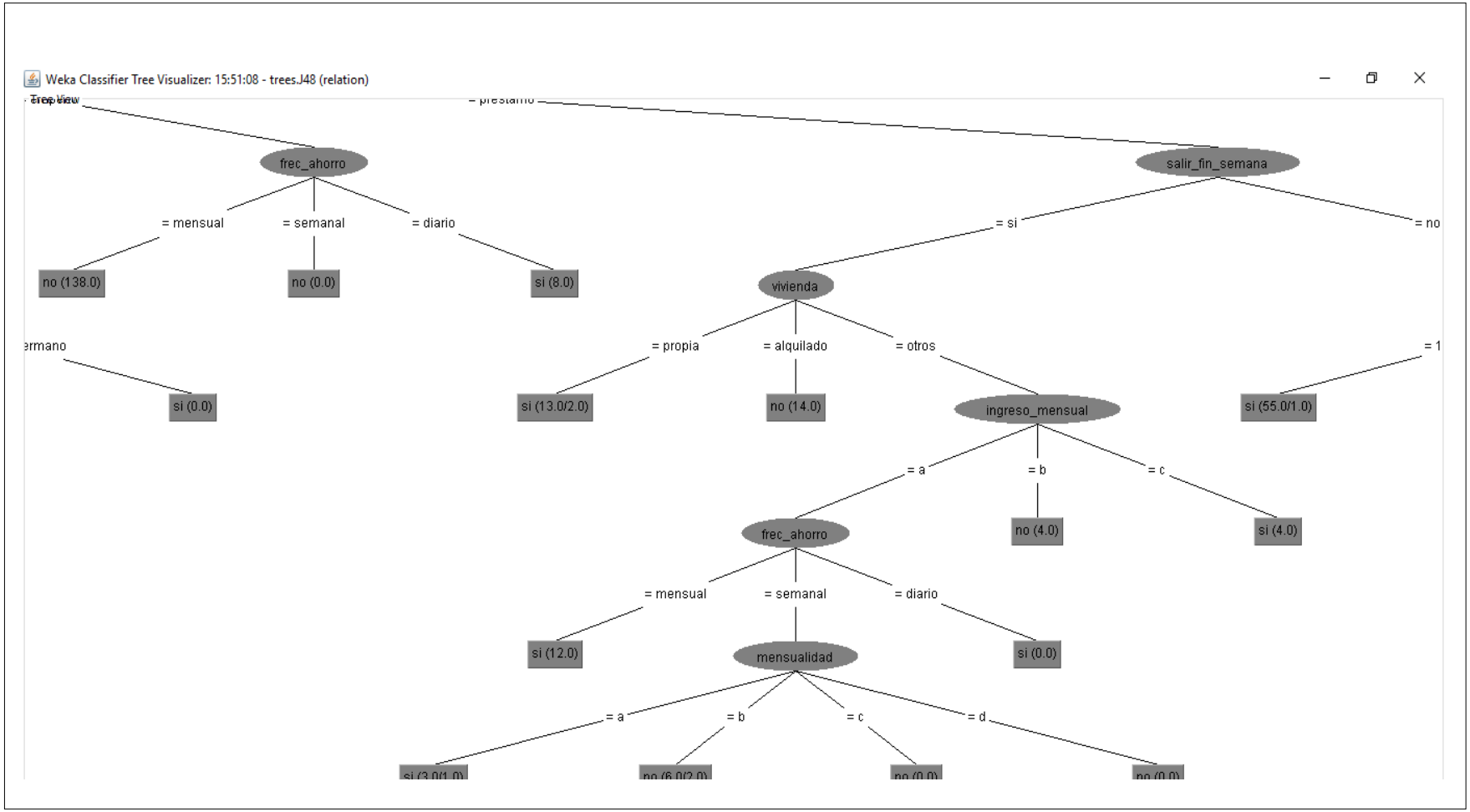


*Figura 34*. Entorno Árbol de decisión Modelo 5

## Modelo 6

## Atributo vivienda

```
Classifier output
J48 pruned tree
------------------

prestamo = aval
|   frec_ahorro = mensual
|   |   ingreso_mensual = a: propia (124.0/6.0)
|   |   ingreso_mensual = b
|   |   |   salir_fin_semana = si: propia (48.0/13.0)
|   |   |   salir_fin_semana = no: otros (21.0)
|   |   ingreso_mensual = c
|   |   |   sexo = hombre: propia (70.0/2.0)
|   |   |   sexo = mujer: alquilado (7.0)
|   frec_ahorro = semanal
|   |   ingreso_mensual = a
|   |   |   mensualidad = a: alquilado (20.0)
|   |   |   mensualidad = b
|   |   |   |   hermanos = 1hermano: otros (7.0)
|   |   |   |   hermanos = 2hermano: propia (9.0)
|   |   |   |   hermanos = 3hermano: propia (0.0)
|   |   |   |   hermanos = masde3hermano: alquilado (9.0)
|   |   |   mensualidad = c: propia (12.0)
|   |   |   mensualidad = d: alquilado (0.0)
|   |   ingreso_mensual = b: alquilado (71.0)
|   |   ingreso_mensual = c
|   |   |   mensualidad = a: alquilado (15.0/4.0)
|   |   |   mensualidad = b: propia (3.0)
|   |   |   mensualidad = c: propia (15.0/2.0)
|   |   |   mensualidad = d: propia (21.0/1.0)
|   frec_ahorro = diario
|   |   mensualidad = a: alquilado (4.0/1.0)
|   |   mensualidad = b: propia (0.0)
|   |   mensualidad = c: propia (0.0)
```

*Figura 35*. Entorno Data Modelo 6

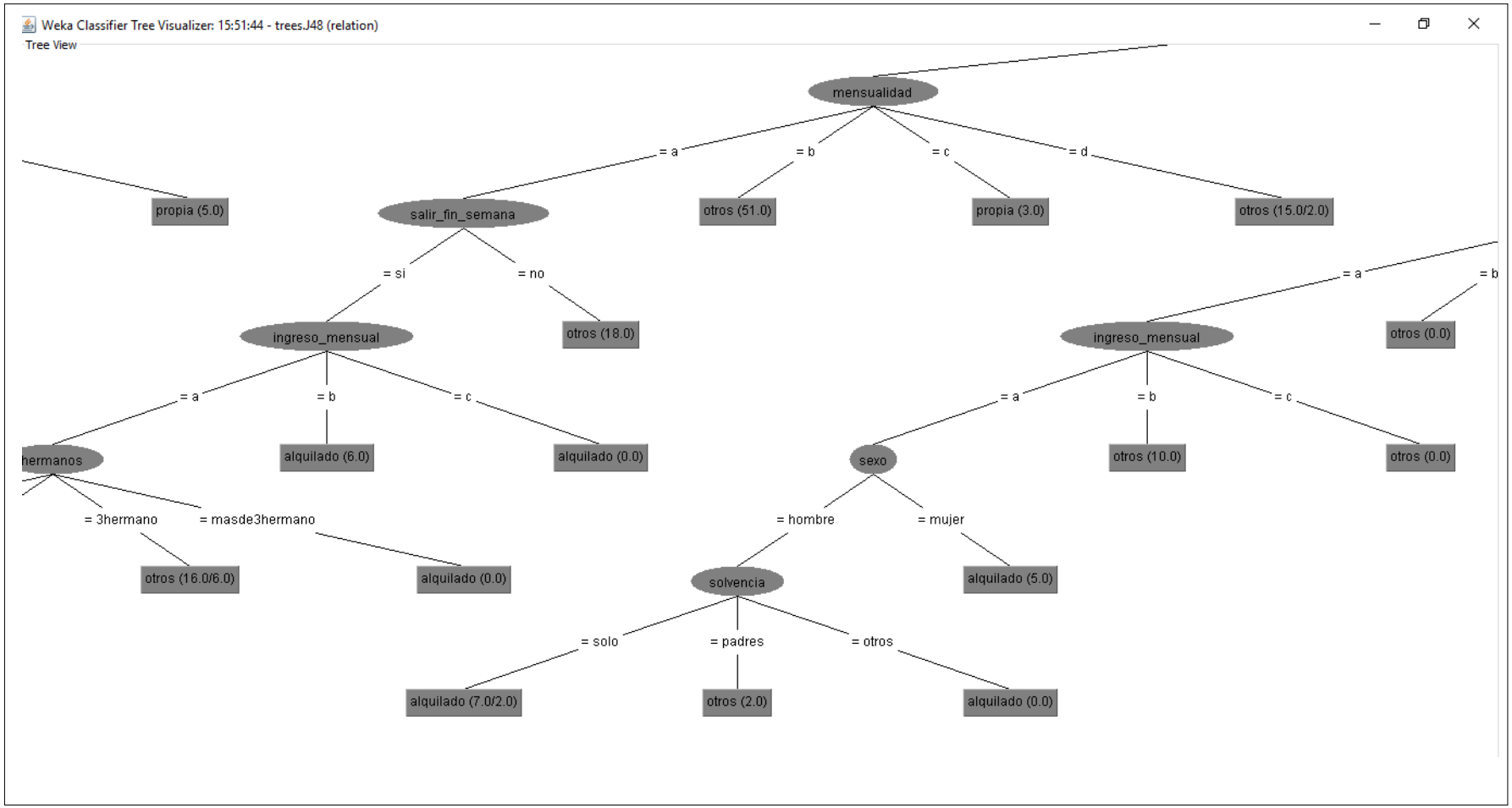


*Figura 36*. Entorno árbol de decisión Modelo 6

## Modelo 7

## Atributo hermanos

Classifier output

```
J48 pruned tree
------------------

frec_ahorro = mensual
|   prestamo = aval
|   |   vivienda = propia
|   |   |   mensualidad = a
|   |   |   |   ingreso_mensual = a
|   |   |   |   |   ayuda_social = si: 3hermano (4.0)
|   |   |   |   |   ayuda_social = no
|   |   |   |   |   |   solvencia = solo: 1hermano (3.0)
|   |   |   |   |   |   solvencia = padres: 1hermano (3.0)
|   |   |   |   |   |   solvencia = otros: 3hermano (2.0)
|   |   |   |   ingreso_mensual = b: 3hermano (4.0)
|   |   |   |   ingreso_mensual = c: 3hermano (30.0/9.0)
|   |   |   mensualidad = b
|   |   |   |   ayuda_social = si: 1hermano (37.0)
|   |   |   |   ayuda_social = no: 2hermano (24.0)
|   |   |   mensualidad = c
|   |   |   |   ingreso_mensual = a
|   |   |   |   |   ayuda_social = si: 3hermano (24.0)
|   |   |   |   |   ayuda_social = no: 1hermano (4.0)
|   |   |   |   ingreso_mensual = b: 3hermano (11.0)
|   |   |   |   ingreso_mensual = c: 2hermano (34.0)
|   |   |   mensualidad = d
|   |   |   |   ayuda_social = si: 1hermano (30.0/6.0)
|   |   |   |   ayuda_social = no
|   |   |   |   |   ingreso_mensual = a: 2hermano (7.0/2.0)
|   |   |   |   |   ingreso_mensual = b: 3hermano (4.0)
|   |   |   |   |   ingreso_mensual = c: 3hermano (0.0)
|   |   vivienda = alquilado: 1hermano (11.0/1.0)
|   |   vivienda = otros: 1hermano (38.0/1.0)
```

*Figura 37*. Entorno Data Modelo 7

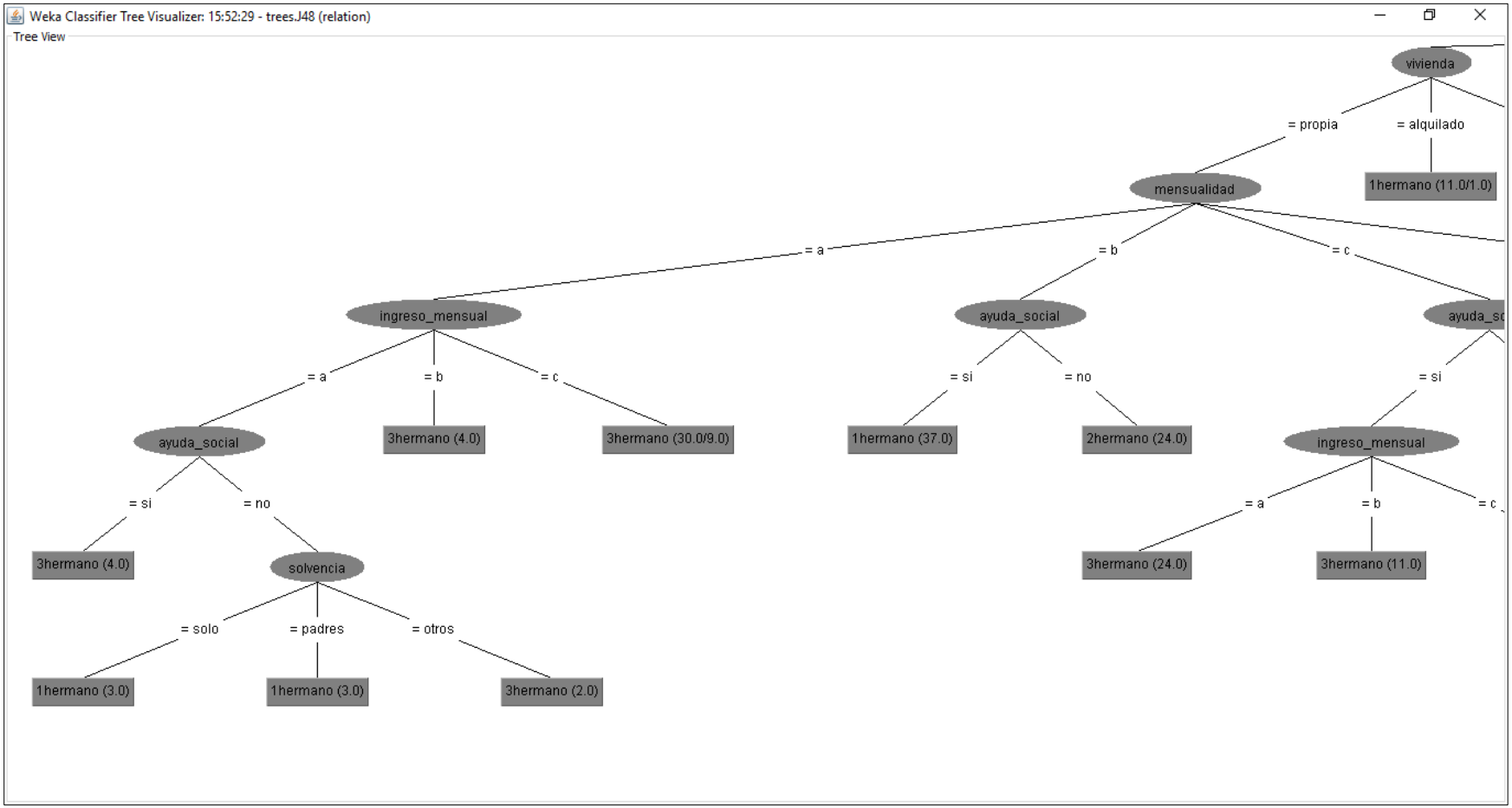


*Figura 38.* Entorno árbol de decisión Modelo 7

## Modelo 8

## Atributo ingresomensual

Classifier output

```
J48 pruned tree
------------------

hermanos = 1hermano
|   frec_ahorro = mensual
|   |   sexo = hombre
|   |   |   mensualidad = a
|   |   |   |   vivienda = propia: a (9.0/1.0)
|   |   |   |   vivienda = alquilado: a (5.0)
|   |   |   |   vivienda = otros
|   |   |   |   |   salir_fin_semana = si: a (6.0/2.0)
|   |   |   |   |   salir_fin_semana = no: b (24.0)
|   |   |   mensualidad = b: a (46.0/4.0)
|   |   |   mensualidad = c: a (28.0)
|   |   |   mensualidad = d
|   |   |   |   prestamo = aval: b (29.0/9.0)
|   |   |   |   prestamo = empeño: b (16.0/6.0)
|   |   |   |   prestamo = prestamo: a (13.0)
|   |   sexo = mujer
|   |   |   mensualidad = a
|   |   |   |   prestamo = aval
|   |   |   |   |   solvencia = solo: b (0.0)
|   |   |   |   |   solvencia = padres: b (4.0)
|   |   |   |   |   solvencia = otros: a (3.0/1.0)
|   |   |   |   prestamo = empeño: a (5.0)
|   |   |   |   prestamo = prestamo
|   |   |   |   |   vivienda = propia: a (6.0/2.0)
|   |   |   |   |   vivienda = alquilado: c (10.0/3.0)
|   |   |   |   |   vivienda = otros: a (4.0)
|   |   |   mensualidad = b
|   |   |   |   salir_fin_semana = si
|   |   |   |   |   prestamo = aval: a (6.0)
```

*Figura 39.* Entorno Data Modelo 8

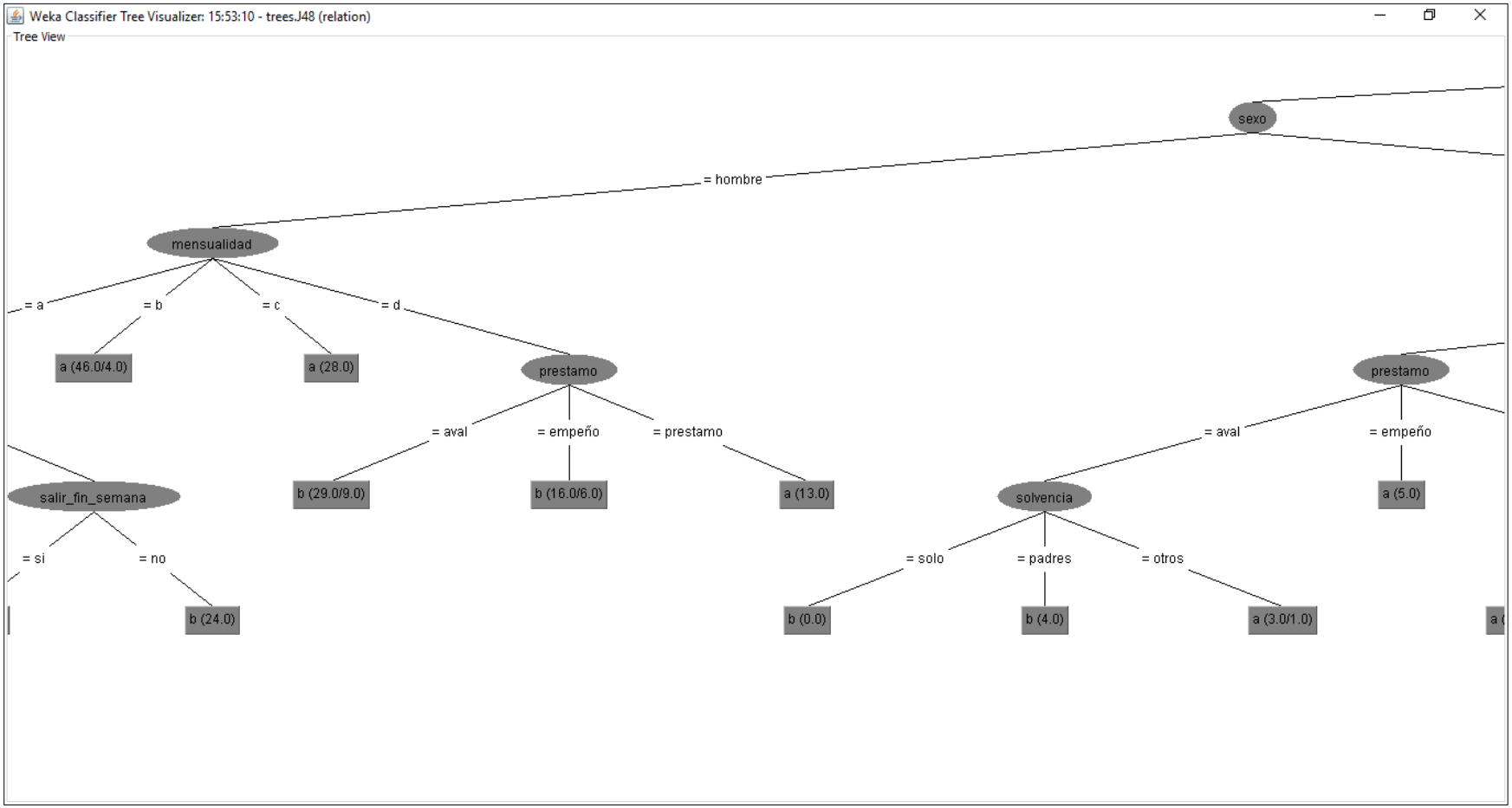


*Figura 40*. Entorno árbol de decisión Modelo 8

## Modelo 9

## Atributo salir fin semana

Classifier output

```
J48 pruned tree
------------------

hermanos = 1hermano
|   frec_ahorro = mensual
|   |   vivienda = propia
|   |   |   ingreso_mensual = a
|   |   |   |   prestamo = aval
|   |   |   |   |   mensualidad = a: si (6.0/2.0)
|   |   |   |   |   mensualidad = b: si (33.0/12.0)
|   |   |   |   |   mensualidad = c: no (4.0)
|   |   |   |   |   mensualidad = d: si (12.0)
|   |   |   |   prestamo = empeño: si (0.0)
|   |   |   |   prestamo = prestamo: no (12.0/3.0)
|   |   |   ingreso_mensual = b: si (12.0)
|   |   |   ingreso_mensual = c
|   |   |   |   prestamo = aval: no (4.0)
|   |   |   |   prestamo = empeño: no (0.0)
|   |   |   |   prestamo = prestamo: si (3.0)
|   |   vivienda = alquilado
|   |   |   mensualidad = a: si (20.0/4.0)
|   |   |   mensualidad = b
|   |   |   |   ayuda_social = si: no (16.0)
|   |   |   |   ayuda_social = no: si (8.0)
|   |   |   mensualidad = c: no (11.0)
|   |   |   mensualidad = d
|   |   |   |   prestamo = aval: si (7.0/1.0)
|   |   |   |   prestamo = empeño: no (16.0)
|   |   |   |   prestamo = prestamo: no (26.0)
|   |   vivienda = otros
|   |   |   mensualidad = a
|   |   |   |   ingreso_mensual = a
```

*Figura 41*. Entorno Data Modelo 9

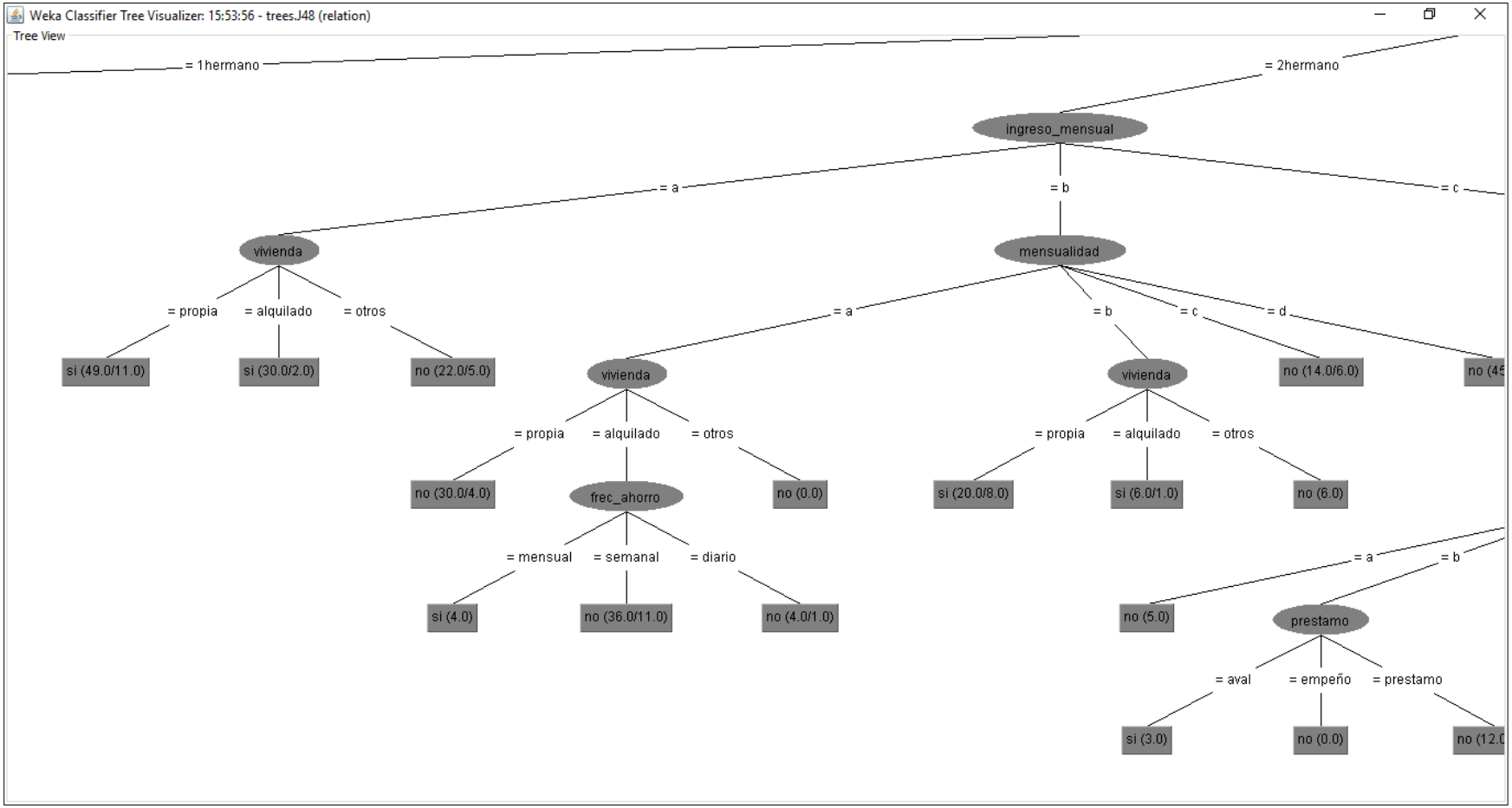


*Figura 42.* Entorno árbol de decisión Modelo 9

## Modelo 10

## Atributo mensualidad

Classifier output

```
ingreso_mensual = a
|   vivienda = propia
|   |   prestamo = aval
|   |   |   hermanos = 1hermano
|   |   |   |   frec_ahorro = mensual
|   |   |   |   |   ayuda_social = si: b (45.0/12.0)
|   |   |   |   |   ayuda_social = no
|   |   |   |   |   |   salir_fin_semana = si: a (4.0)
|   |   |   |   |   |   salir_fin_semana = no: c (6.0/2.0)
|   |   |   |   frec_ahorro = semanal: c (12.0)
|   |   |   |   frec_ahorro = diario: b (0.0)
|   |   |   hermanos = 2hermano: b (38.0/5.0)
|   |   |   hermanos = 3hermano
|   |   |   |   ayuda_social = si: c (30.0/6.0)
|   |   |   |   ayuda_social = no: a (4.0/2.0)
|   |   |   hermanos = masde3hermano: b (0.0)
|   |   prestamo = empeño
|   |   |   frec_ahorro = mensual: a (10.0)
|   |   |   frec_ahorro = semanal: c (3.0)
|   |   |   frec_ahorro = diario: a (0.0)
|   |   prestamo = prestamo
|   |   |   hermanos = 1hermano
|   |   |   |   salir_fin_semana = si: a (3.0)
|   |   |   |   salir_fin_semana = no: d (9.0/3.0)
|   |   |   hermanos = 2hermano: d (0.0)
|   |   |   hermanos = 3hermano: d (0.0)
|   |   |   hermanos = masde3hermano
|   |   |   |   salir_fin_semana = si: c (10.0/3.0)
|   |   |   |   salir_fin_semana = no: b (9.0/5.0)
|   vivienda = alquilado
|   |   salir_fin_semana = si
|   |   |   prestamo = aval
|   |   |   |   hermanos = 1hermano: a (10.0)
```

*Figura 43.* Entorno árbol de decisión Modelo 10

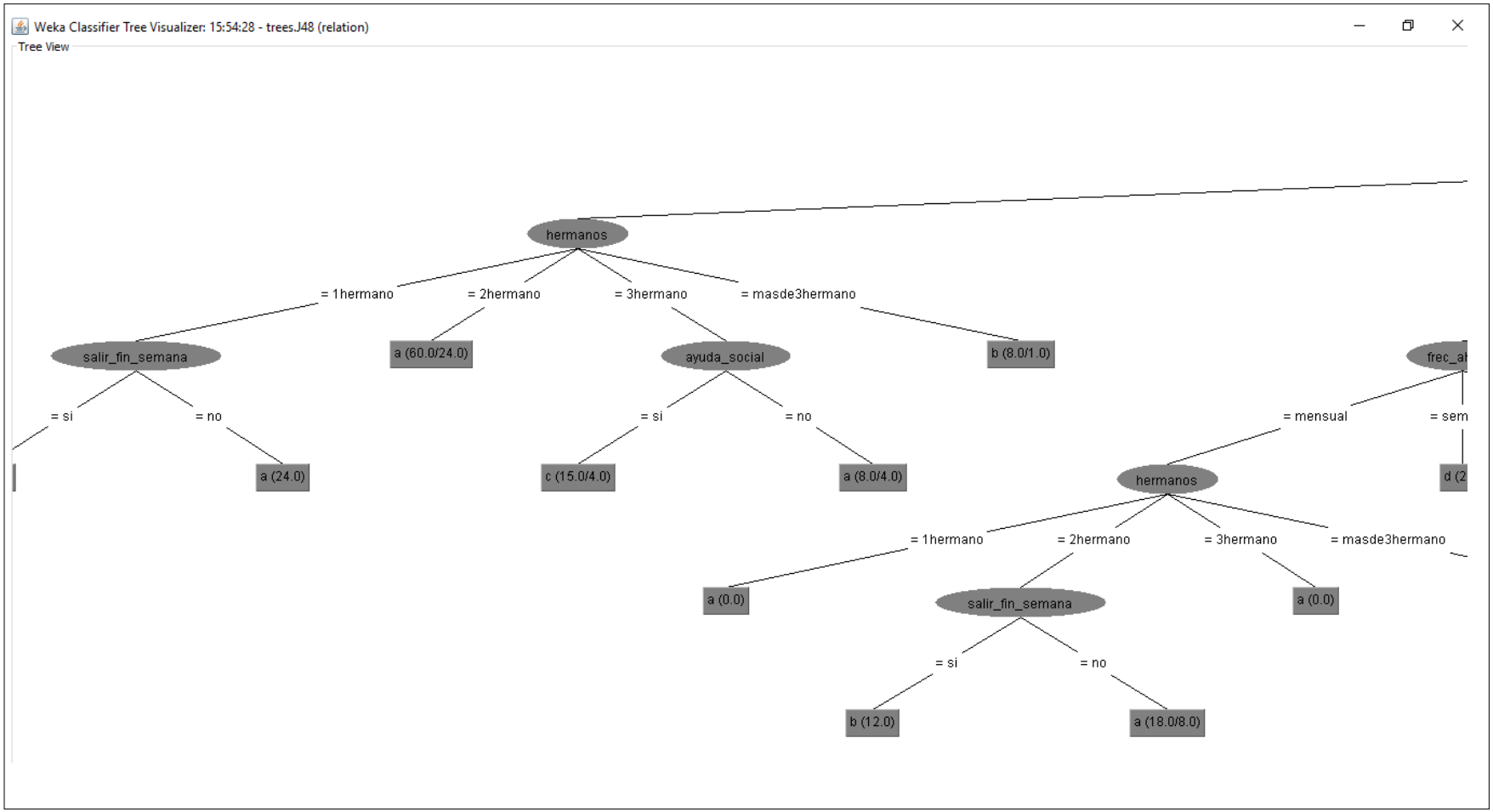


*Figura 44*. Entorno árbol de decisión Modelo 10

## Descripción del modelo

A continuación, vamos a describir el resultado de la ejecución de cada uno de los modelos para cada objetivo, estos resultados se estudiarán más a fondo en la etapa de evaluación.

Al evaluar cada modelo se obtuvo lo siguiente.

## Modelo 1

```
Correctly Classified Instances         639               52.9412 %
Incorrectly Classified Instances       568               47.0588 %
Kappa statistic                          0.2696
Mean absolute error                      0.3699
Root mean squared error                  0.4301
Relative absolute error                 84.9087 %
Root relative squared error             92.1482 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.814     0.528      0.497     0.814     0.617      0.706    solo
                 0.256     0.085      0.637     0.256     0.365      0.68     padres
                 0.486     0.125      0.551     0.486     0.516      0.771    otros
Weighted Avg.    0.529     0.268      0.562     0.529     0.5        0.712
```

*Figura 45.* Entorno de evaluación 1

## Modelo 2

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1169               96.8517 %
Incorrectly Classified Instances        38                3.1483 %
Kappa statistic                          0.9473
Mean absolute error                      0.0359
Root mean squared error                  0.134
Relative absolute error                  9.0202 %
Root relative squared error             30.0379 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.973     0.018      0.984     0.973     0.979      0.996    mensual
                 0.976     0.031      0.934     0.976     0.954      0.992    semanal
                 0.938     0.002      0.989     0.938     0.963      0.996    diario
Weighted Avg.    0.969     0.019      0.969     0.969     0.969      0.995
```

*Figura 46.* Entorno de evaluación 2

## Modelo 3

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1101               91.2179 %
Incorrectly Classified Instances       106                8.7821 %
Kappa statistic                          0.8233
Mean absolute error                      0.1352
Root mean squared error                  0.26
Relative absolute error                 27.1725 %
Root relative squared error             52.1275 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.898    0.076      0.911     0.898     0.905      0.968    hombre
                 0.924    0.102      0.913     0.924     0.919      0.968    mujer
Weighted Avg.    0.912    0.09       0.912     0.912     0.912      0.968
```

*Figura 47*. Entorno de evaluación 3

## Modelo 4

```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1164               96.4374 %
Incorrectly Classified Instances        43                3.5626 %
Kappa statistic                          0.945
Mean absolute error                      0.0409
Root mean squared error                  0.143
Relative absolute error                  9.4511 %
Root relative squared error             30.7437 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.954    0.001      0.998     0.954     0.976      0.994    aval
                 0.938    0.01       0.966     0.938     0.952      0.995    empeño
                 0.989    0.045      0.934     0.989     0.961      0.991    prestamo
Weighted Avg.    0.964    0.02       0.966     0.964     0.964      0.993
```

*Figura 48*. Entorno de evaluación 4

## Modelo 5

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1169               96.8517 %
Incorrectly Classified Instances        38                3.1483 %
Kappa statistic                          0.9136
Mean absolute error                      0.0504
Root mean squared error                  0.1587
Relative absolute error                 13.4489 %
Root relative squared error             36.6829 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.992     0.103      0.967     0.992     0.979      0.991    si
                 0.897     0.008      0.975     0.897     0.934      0.991    no
Weighted Avg.    0.969     0.079      0.969     0.969     0.968      0.991
```

*Figura 49*. Entorno de evaluación 5

## Modelo 6

```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1119               92.7092 %
Incorrectly Classified Instances        88                7.2908 %
Kappa statistic                          0.8898
Mean absolute error                      0.0761
Root mean squared error                  0.1951
Relative absolute error                 17.2402 %
Root relative squared error             41.5217 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.937     0.047      0.925     0.937     0.931      0.985    propia
                 0.921     0.032      0.918     0.921     0.92       0.989    alquilado
                 0.921     0.031      0.938     0.921     0.929      0.989    otros
Weighted Avg.    0.927     0.037      0.927     0.927     0.927      0.988
```

*Figura 50*. Entorno de evaluación 6

## Modelo 7

```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1081               89.5609 %
Incorrectly Classified Instances       126               10.4391 %
Kappa statistic                          0.8519
Mean absolute error                      0.0815
Root mean squared error                  0.2019
Relative absolute error                 23.0031 %
Root relative squared error             47.9661 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.92      0.06       0.898     0.92      0.909      0.976    1hermano
                 0.895     0.066      0.874     0.895     0.884      0.969    2hermano
                 0.918     0.022      0.891     0.918     0.904      0.992    3hermano
                 0.806     0.005      0.964     0.806     0.878      0.99     masde3hermano
Weighted Avg.    0.896     0.048      0.897     0.896     0.895      0.978
```

*Figura 51.* Entorno de evaluación 7

## Modelo 8

```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1094               90.6379 %
Incorrectly Classified Instances       113                9.3621 %
Kappa statistic                          0.8481
Mean absolute error                      0.0918
Root mean squared error                  0.2142
Relative absolute error                 22.143  %
Root relative squared error             47.0606 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.94      0.093      0.908     0.94      0.923      0.977    a
                 0.876     0.056      0.874     0.876     0.875      0.974    b
                 0.87      0.009      0.959     0.87      0.912      0.993    c
Weighted Avg.    0.906     0.065      0.907     0.906     0.906      0.979
```

*Figura 52.* Entorno de evaluación 8

## Modelo 9

```
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1033               85.5841 %
Incorrectly Classified Instances       174               14.4159 %
Kappa statistic                          0.7034
Mean absolute error                      0.2013
Root mean squared error                  0.3172
Relative absolute error                 40.9135 %
Root relative squared error             63.9645 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.913     0.218      0.844     0.913     0.877      0.934    si
                 0.782     0.087      0.875     0.782     0.826      0.934    no
Weighted Avg.    0.856     0.161      0.857     0.856     0.855      0.934
```

*Figura 53.* Entorno de evaluación 9

## Modelo 10

```
Time taken to build model: 0.04 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         958               79.3703 %
Incorrectly Classified Instances       249               20.6297 %
Kappa statistic                          0.714
Mean absolute error                      0.1381
Root mean squared error                  0.2628
Relative absolute error                 38.3082 %
Root relative squared error             61.8977 %
Total Number of Instances             1207

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.827     0.109      0.812     0.827     0.819      0.951    a
                 0.754     0.079      0.803     0.754     0.778      0.945    b
                 0.935     0.06       0.775     0.935     0.848      0.981    c
                 0.63      0.038      0.753     0.63      0.686      0.954    d
Weighted Avg.    0.794     0.08       0.793     0.794     0.791      0.955
```

*Figura 54*. Entorno de evaluación 10

### 3.5.3 Evaluar el Modelo

Aunque en el paso 5 de la metodología CRISP-DM (evaluación) también se haga una evaluación de los modelos generados, en este apartado dicha evaluación está más orientada a los objetivos de la minería de datos, mientras que en el siguiente paso de la metodología, la evaluación se orienta más a los objetivos de negocio, si bien ambos objetivos están muy relacionados entre sí en este proyecto.

En términos de minería de datos, una buena manera de evaluar la efectividad de los modelos es utilizando los dos indicadores que se establecieron en el plan de pruebas de este documento, dichos indicadores son el error cuadrático medio (root mean squared error) y el error absoluto medio (mean absolute error).

Muy aparte de los valores ya descritos también se utilizará para evaluar la matriz de confusión ya descrita con anterioridad.

Según cada modelo se mostrará la matriz de confusión correspondiente:

**Modelo 1**

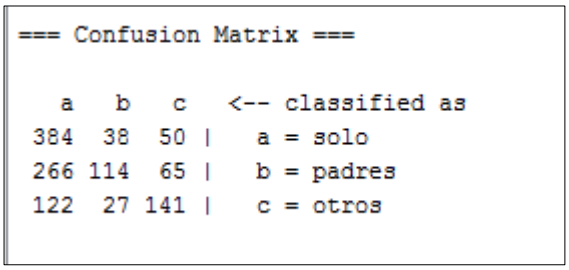


*Figura 55*. Matriz de confusión 1

Para este modelo se denota que la diagonal empezando en el numero 384 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso. Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 84.9087 %

Root relative squared error 92.1482 %

**Modelo 2**

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 622  15   2 |   a = mensual
   9 367   0 |   b = semanal
   1  11 180 |   c = diario
```

*Figura 56.* Matriz de confusión 2

Para este modelo se denota que la diagonal empezando en el numero 622 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso.

Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 9.0202 %

Root relative squared error 30.0379 %

**Modelo 3**

```
=== Confusion Matrix ===

   a   b   <-- classified as
 503  57 |   a = hombre
  49 598 |   b = mujer
```

*Figura 57.* Matriz de confusión 3

Para este modelo se denota que la diagonal empezando en el numero 503 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso. Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 27.1725 %

Root relative squared error 52.1275 %

**Modelo 4**

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 440   5  16 |   a = aval
   0 256  17 |   b = empeño
   1   4 468 |   c = prestamo
```

*Figura 58*. Matriz de confusión 4

Para este modelo se denota que la diagonal empezando en el numero 440 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso.

Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 9.4511 %

Root relative squared error 30.7437 %

**Modelo 5**

```
=== Confusion Matrix ===

   a   b   <-- classified as
 899   7 |   a = si
  31 270 |   b = no
```

*Figura 59.* Matriz de confusión 5

Para este modelo se denota que la diagonal empezando en el numero 899 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso. Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 13.4489 %

Root relative squared error 36.6829 %

**Modelo 6**

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 429  15  14 |   a = propia
  16 315  11 |   b = alquilado
  19  13 375 |   c = otros
```

*Figura 60*. Matriz de confusión 6

Para este modelo se denota que la diagonal empezando en el numero 429 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso.

Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 17.2402 %

Root relative squared error 41.5217 %

**Modelo 7**

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 403  33   1   1 |   a = 1hermano
  18 366  21   4 |   b = 2hermano
   8   8 179   0 |   c = 3hermano
  20  12   0 133 |   d = masde3hermano
```

*Figura 61*. Matriz de confusión 7

Para este modelo se denota que la diagonal empezando en el numero 403 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso. Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 23.0031 %

Root relative squared error 47.9661 %

**Modelo 8**

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 561  34   2 |   a = a
  39 325   7 |   b = b
  18  13 208 |   c = c
```

*Figura 62*. Matriz de confusión 8

Para este modelo se denota que la diagonal empezando en el numero 561 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso. Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 22.143 %

Root relative squared error 47.0606 %

**Modelo 9**

```
=== Confusion Matrix ===

   a   b   <-- classified as
 621  59 |   a = si
 115 412 |   b = no
```

*Figura 63*. Matriz de confusión 9

Para este modelo se denota que la diagonal empezando en el numero 621 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso. Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 40.9135 %

Root relative squared error 63.9645 %

**Modelo 10**

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 363  45  27   4 |   a = a
  50 273  12  27 |   b = b
   4   2 203   8 |   c = c
  30  20  20 119 |   d = d
```

*Figura 64*. Matriz de confusión 10

Para este modelo se denota que la diagonal empezando en el nuero 363 es mayor esto nos da un alto indicio de probabilidad de que el modelo está prediciendo según el caso. Muy aparte del porcentaje que se tiene al hacer la validación cruzada:

Relative absolute error 38.3082 %

Root relative squared error 61.8977 %

## 3.6 Evaluación

En esta fase de la metodología se intentan evaluar los modelos generados, pero en esta ocasión la evaluación se hace desde el punto de vista de los objetivos de negocio en lugar de los objetivos de minería de datos. Una vez realizada esta evaluación, se debe decidir si los objetivos han sido cumplidos y de ser así se puede avanzar a la fase de implantación, de lo contrario se tendría que identificar cualquier factor que se haya podido pasar por alto y hacer una revisión del proceso.

### 3.6.1 Evaluar los Resultados

Desde el punto de vista del negocio, se había establecido como criterio de éxito principal el poder realizar predicciones con un porcentaje de fiabilidad

"aceptable", este criterio puede ser algo subjetivo, por lo que es inevitable apoyarse principalmente en los criterios de éxito desde el punto de vista de la minería de datos que son mucho más específicos y precisos. Además, para poder calificar como aceptable o no las predicciones que se van a realizar es necesario tener una base objetiva, como lo son los indicadores estadísticos que se han obtenido al ejecutar los modelos. También sería conveniente la evaluación de los resultados por parte de un grupo de expertos en la minería de datos, si se contara con ellos. En cualquier caso, basándonos en los indicadores obtenidos mediante la herramienta de minería de datos, a continuación, podemos hacer una evaluación de cada modelo para así descartar aquel que no cumpla con unos requisitos mínimos.

**Modelo 1**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 52.9412 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 2**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 96.8517 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 3**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 91.2179 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 4**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 96.4374 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 5**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 96.8517 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 6**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 92.7092 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 7**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 89.5609 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 8**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 90.6379 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 9**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 85.5841 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelo 10**

Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 79.3703 %, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

**Modelos aprobados**

Por las razones explicadas los modelos aprobados son el modelo 2 al modelo 10 que cumplen con los criterios de éxito de negocio, mientras que el modelo 1 será descartado por no cumplir con los requisitos de negocio ni de minería de datos.

**Simulación de los algoritmos con un sistema.**

Al ya tener los algoritmos generados pasamos a interpretarlos en una interfaz para que interactúen con el usuario final, se valida el entrenamiento con los últimos datos de confianza recogidos para luego pasarlos a código visual Basic, de esta manera el usuario pueda hacer las predicciones sin ningún percance.

A continuación, se mostrará el cuadro de la probabilidad de acierto del sistema, junto a la interfaz de prueba que se conectará con los datos entrenados para dar la predicción lo más asertiva posible.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

sexo = hombre: no (560.0)
sexo = mujer: si (647.0/3.0)

Number of Leaves  :     2

Size of the tree :      3


Time taken to build model: 0.04 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1204               99.7514 %
Incorrectly Classified Instances         3                0.2486 %
Kappa statistic                          0.995
Mean absolute error                      0.0049
Root mean squared error                  0.0497
Relative absolute error                  0.9941 %
Root relative squared error              9.9703 %
Total Number of Instances             1207
```

*Figura 65*. Algoritmo de predicción

Obteniendo un 99,8% de exactitud de predicción se prosigue a interactuar con el sistema creado para mejor entendimiento.

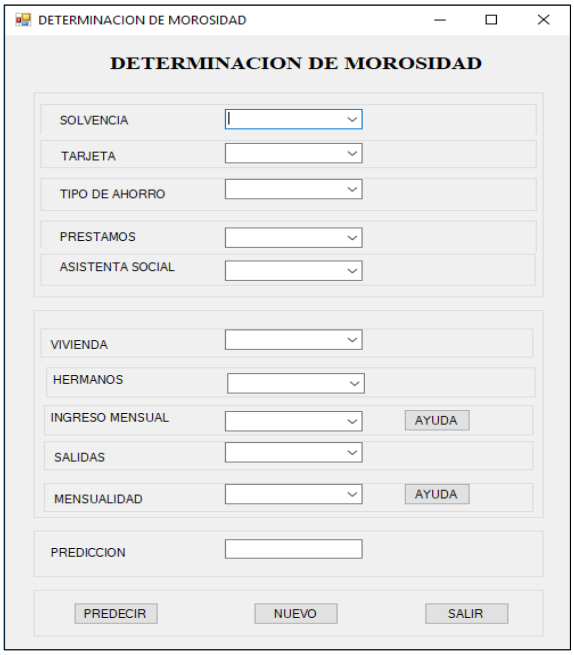


*Figura 66*. Interacción con el usuario

Se usó código visual basic.net para crear la infernas comando e interacción con los modelos predictivos.

```
    PREDICCION.Text = "NO MOROSO"
ElseIf SOLVENCIA.SelectedItem = "PADRES" And TARJETA.SelectedItem = "SI" And TIP_AHORRO.SelectedItem = "MENSUAL" A
    PREDICCION.Text = "NO MOROSO"
ElseIf SOLVENCIA.SelectedItem = "OTROS" And TARJETA.SelectedItem = "NO" And TIP_AHORRO.SelectedItem = "DIARIO" And
    PREDICCION.Text = "NO MOROSO"
ElseIf SOLVENCIA.SelectedItem = "PADRES" And TARJETA.SelectedItem = "SI" And TIP_AHORRO.SelectedItem = "MENSUAL" A
    PREDICCION.Text = "NO MOROSO"
ElseIf SOLVENCIA.SelectedItem = "OTROS" And TARJETA.SelectedItem = "NO" And TIP_AHORRO.SelectedItem = "DIARIO" And
    PREDICCION.Text = "MOROSO"


End If

minum = 1 + Int(Rnd() * (10 - 1 + 1))
```

*Figura 67.* Codificación de estructura

Mediante una serie de interpretaciones de código weka se van desglosando los algoritmos que se usan para que el interfaz funcione y de esta manera se pueda mejorar la interacción con el usuario final.

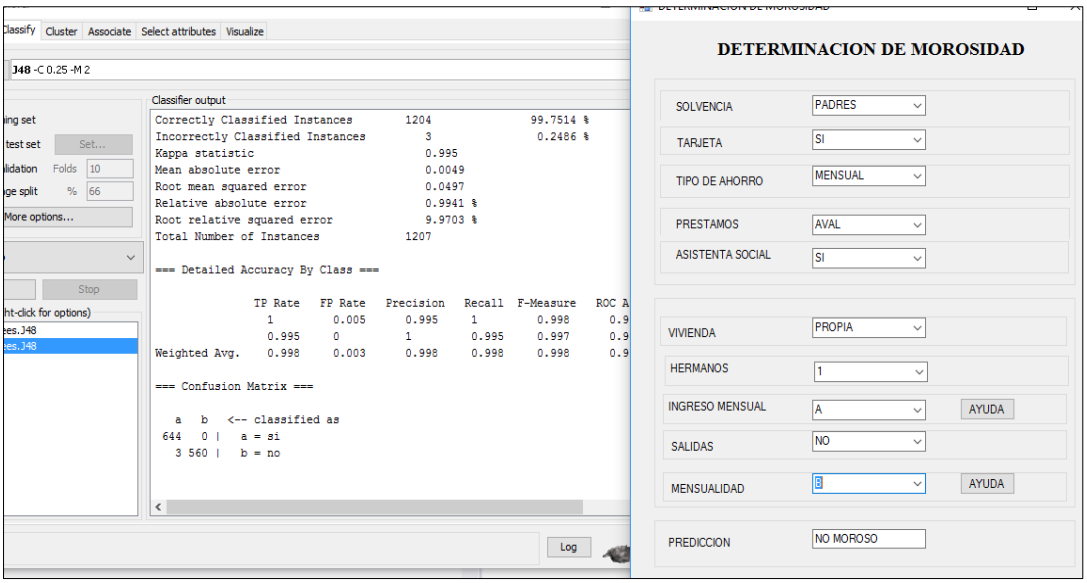


*Figura 68.* Interacción con el usuario 2

### 3.6.2 Revisar el Proceso

El proceso hasta este punto se ha ejecutado tal y como estaba previsto, si bien ha habido complicaciones a la hora de realizar el modelo 1, ya que se han obtenido valores muy deficientes para la confianza predictiva, el error absoluto medio y el error cuadrático medio. La causa de estos malos valores posiblemente se encuentre en que no disponemos de todos los datos que se podrían necesitar para hacer una

predicción fiable sobre el modelo descrito. En cualquier caso, para el presente proyecto se ha decidido descartar este objetivo.

### 3.6.3 Determinar los Próximos Pasos

El siguiente paso a realizar en el proyecto es el de ejecutar la etapa de implantación para los modelos 2 al 10.

## 3.7 Implantación

Esta es la última fase de la metodología CRISP-DM y el objetivo de la misma es el de explicar al cliente como poner en funcionamiento el proyecto que se ha construido en las fases anteriores, así como exponer los resultados obtenidos al cliente de forma que lo pueda entender fácilmente. Otro objetivo de esta fase es el de crear una estrategia para el mantenimiento del proyecto y producir un informe en el que se incluyan posibles mejoras para el futuro y un listado de las dificultades encontradas a la hora de realizarlo

### 3.7.1 Planear la Implantación

Para poder implantar este proyecto en el negocio sería necesario en primer lugar tener acceso a la base de datos total del negocio, es decir la base de datos que contiene toda la información relativa a los alumnos de la universidad. A partir de ahí, los pasos a seguir serían los mismos que se han seguido en este documento desde la comprensión del negocio hasta la implantación. Si bien, cabe decir que habrá algunas fases, como la de comprensión y preparación de los datos, que en el negocio real probablemente sean más complejas y llevarán más tiempo que en este proyecto ya que se puede esperar que en la base de datos real se tengan muchos más registros y estos mismos contengan más ruido que en nuestra base de datos ficticia creada específicamente para este uso.

### 3.7.2 Planear la Monitorización y Mantenimiento

La supervisión y mantenimiento de la implementación del presente proyecto es una fase importante del mismo debido a que los datos que se procesan con mucha frecuencia pueden ser modificados por el personal de la universidad. Los datos pueden ser modificados por diferentes motivos como haber realizado una codificación incorrecta, haber asignado una nota incorrecta al alumno, etc. El volumen de estos datos en movimiento es grande motivo por el cual la extracción de las muestras debe ser realizada cuidadosamente y realizando siempre backups de los

datos explotados en cada proceso. La minería de datos debería ser realizada en periodos de cuatro meses (cuatrimestres) ya que esta es la medida de tiempo utilizada en la universidad para realizar los exámenes y asignar las notas finales a los alumnos, sin embargo esta medida podría variar en cualquier momento en función del plan de estudios que esté vigente en cada momento.

Como plan de supervisión y mantenimiento se podría establecer los siguientes procesos:

• Extracción y almacenamiento cuatrimestral de los datos guardando la información obtenida en formato de hoja de cálculo

• Distribución de los datos en función de los modelos de software de minería de datos a trabajar.

• Los archivos de la explotación de datos deberán ser guardados en soporte magnético en la propia universidad, almacenándolos por ejemplo en carpetas ordenadas por procesos cuatrimestrales.

• Los resultados obtenidos en cada explotación de datos deberán ser llevados a formato de hoja de cálculo y generar gráficas de distintos tipos para una mejor visualización e interpretación de los resultados obtenidos en cada periodo.

### 3.7.3 Producir el Informe Final

En este paso se debe presentar un informe resumiendo los puntos importantes del proyecto y la experiencia adquirida durante su desarrollo. El público al que va dirigido este informe sería el personal de la universidad encargado de la docencia (profesores, directores de departamento, etc.) de tal manera que se pueda estudiar la situación actual y tomar medidas correctivas para la mejora del servicio académico. Cabe decir que parte de este informe final será presentado de manera oral con una presentación, por lo que en este apartado solamente haremos un breve resumen.

El uso de la metodología CRISP-DM en este proyecto ha permitido encontrar un comportamiento predictivo a la hora de estimar la duración de la carrera de los alumnos y la nota media de los mismos. Se ha podido encontrar un plan de extracción, normalización, y codificación de datos para la realización de procesos de minería de datos cuatrimestrales.

De los tres objetivos de minería de datos iniciales que se habían fijado se han podido alcanzar dos de ellos (modelos 2 al 10). Además, al margen de estos objetivos, se han sacado otras conclusiones a partir de los datos estudiados, concretamente se han identificado las asignaturas más problemáticas para los alumnos de cada una de las titulaciones estudiadas.

Repasando las diferentes etapas que hemos seguido para llegar al objetivo:

La primera etapa ha sido una de las más laboriosas por no tener una base de datos de la que partir. Esto ha supuesto que tengamos que generar nosotros mismos un conjunto de datos sobre el que trabajar. Para poder hacer una simulación lo más real posible, no valía con generar datos aleatorios, si no que se ha tenido que desarrollar un pequeño programa en Java que generase estos datos de manera automática, debido a la gran cantidad de datos que necesitábamos manejar para hacer una estimación lo más precisa posible.

Cuando ya disponíamos de la base de datos sobre la que ejercer la minería de datos, se hizo un análisis de la estructura de los datos y la información contenida.

A continuación, se realizó la elección de las técnicas de modelado y la ejecución de dichas técnicas sobre los datos empleando la herramienta escogida para ello (Weka). Esta herramienta facilitó por completo la aplicación de los modelos ya que nos permitió ver de manera muy intuitiva y visual cuales eran las técnicas más adecuadas para nuestra base de datos.

Por último, una vez obtenidos los modelos, se analizaron para determinar la adecuación o no de los mismos. En este caso determinamos que los modelos 2 al modelo 10 podrían ser válidos para nuestros objetivos y se descartó el 1 por no ser lo suficientemente fiable.

Realizados todos estos pasos se presentan los resultados alcanzados al público que es el objetivo de este apartado.

### 3.7.4 Revisar el Proyecto

En esta última etapa de la metodología se debe hacer una evaluación de aquellas cosas que se hicieron correctamente y aquellas que no, así como posibles mejoras para que en las futuras ejecuciones de la minería de datos se vayan puliendo los fallos y se obtengan mejores resultados.

En primer lugar, y como ya se ha comentado anteriormente en otros apartados, el mayor lastre que se ha ido arrastrando a lo largo de este proyecto es el de no disponer de una base de datos real sobre la que actuar ya que esto condiciona en gran medida los resultados obtenidos. A pesar de haber intentado generar unos datos lo más veraces posible, no cabe duda que existen multitud de factores que no podemos manejar y que disponer de los datos reales con las notas y demás características de los alumnos incrementaría aún más la fiabilidad de los modelos de minería de datos elegidos en este proyecto. Esto se puede interpretar como algo positivo ya que, si hemos dado por válidos dos de los tres modelos empleados, y sin disponer de la cantidad y veracidad de los datos que se manejan en la base de datos de la universidad, esto quiere decir que si el proyecto saliera en real los resultados mejorarían aún más

# CAPÍTULO IV
# ANÁLISIS DE RESULTADOS Y CONTRASTACIÓN DE LA HIPÓTESIS

## 4.1 POBLACIÓN Y MUESTRA

### 4.1.1 Población

Se considera como unidad de medida al realizar un modelo predictivo en la Universidad Autónoma del Perú. Por la naturaleza del comportamiento de los estudiantes al solicitar un refinanciamiento de deuda, resulta adecuado considerar una **población indeterminada**, significa que el registro de refinanciamiento va a ocurrir en cantidades no conocidas**.**

### 4.1.2 Muestra

Conociendo o no el tamaño de la población; una muestra de valor 1200, es un valor adecuado y estándar que se utiliza en varios procesos de investigación. Por lo tanto el tamaño de la muestra:

n = 30 refinanciamiento de deudas

## 4.2 Nivel de confianza

El nivel de confianza será del 95% de la inexperiencia de los investigadores.

La significancia será de 5%.

## 4.3 ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

### 4.3.1 Resultados Genéricos

**Fase: Inicio**

- Modelado de Negocio.
- Antecedentes de la empresa.
- Estructura de la empresa.
- Descripción de productos, servicios y clientes.
- Stakeholders de la empresa.
- Identificación del proceso en la cadena de valor.
- Visión del proyecto.
- Alcance del proyecto.
- Planificación.

**Fase: Elaboración**

- Definición de requerimientos.
- Diagrama de Tablas.
- Explicacion de campos.
- Conexión de datos.
- Levantamiento de Informacion.
- Estudio de datos.
- Mineria de datos.
- Generacion de modelos(beta).

**Fase: Construcción**

- Elaboración de conexión weka
- Elaboración de prototipos del modelo predictivo.
- Cuadros de datos, elaboración de árbol de decisión.

**Fase: Transición**

- Elaboración de Pruebas (Versión Beta).
- Informe final.

## 4.3.2 Resultados Específicos

Tabla 20
*Descripción de Datos*

| | KP1: Tiempo para verificar datos del alumno | | KPI2: Tiempo para verificar alumno moroso | | KPI3: Índice de Riesgo | | KPI4: Tiempo para Determinar el riesgo de morosidad de un estudiante | | KPI5: Tiempo para predecir que alumnos incurrirán en morosidad | | KPI6: Nivel de satisfaccion | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **NUMERO** | **PRE-PRUEBA** | **POST-PRUEBA** | **PRE-PRUEBA** | **POST-PRUEBA** | **PRE-PRUEBA** | **POST-PRUEBA** | **PRE-PRUEBA** | **POST-PRUEBA** | **PRE-PRUEBA** | **POST-PRUEBA** | **PRE-PRUEBA** | **POST-PRUEBA** |
| **1** | 8 | 2 | 8 | 2 | 15 | 5 | 8 | 4 | 20 | 3 | bueno | excelente |
| **2** | 9 | 2 | 9 | 3 | 18 | 6 | 9 | 5 | 25 | 2 | regular | bueno |
| **3** | 8 | 3 | 8 | 4 | 15 | 5 | 8 | 3 | 30 | 3 | regular | bueno |
| **4** | 7 | 4 | 7 | 2 | 15 | 4 | 7 | 3 | 26 | 5 | bueno | excelente |
| **5** | 6 | 2 | \| | 1 | 17 | 5 | 6 | 2 | 25 | 2 | malo | bueno |
| **6** | 10 | 4 | 10 | 2 | 18 | 7 | 10 | 3 | 20 | 2 | malo | regular |
| **7** | 6 | 3 | 6 | 3 | 20 | 8 | 6 | 5 | 30 | 3 | bueno | excelente |
| **8** | 7 | 5 | 7 | 4 | 20 | 6 | 7 | 2 | 28 | 4 | regular | bueno |
| **9** | 6 | 3 | 6 | 5 | 15 | 5 | 6 | 2 | 25 | 2 | regular | bueno |
| **10** | 5 | 4 | 5 | 3 | 14 | 4 | 5 | 3 | 20 | 4 | bueno | excelente |
| **11** | 6 | 5 | 7 | 3 | 16 | 3 | 5 | 4 | 20 | 4 | regular | excelente |
| **12** | 8 | 4 | 5 | 4 | 18 | 5 | 8 | 2 | 20 | 2 | malo | bueno |
| **13** | 7 | 3 | 8 | 5 | 19 | 3 | 8 | 4 | 25 | 1 | regular | bueno |
| **14** | 6 | 2 | 8 | 2 | 20 | 5 | 8 | 3 | 25 | 2 | malo | bueno |
| **15** | 7 | 3 | 8 | 3 | 25 | 6 | 9 | 5 | 20 | 3 | regular | bueno |
| **16** | 5 | 2 | 9 | 4 | 25 | 5 | 10 | 3 | 30 | 4 | bueno | excelente |
| **17** | 8 | 3 | 10 | 5 | 16 | 6 | 9 | 4 | 25 | 5 | bueno | excelente |
| **18** | 8 | 2 | 9 | 5 | 18 | 5 | 8 | 5 | 20 | 3 | bueno | excelente |
| **19** | 8 | 1 | 8 | 3 | 16 | 6 | 7 | 4 | 20 | 8 | bueno | excelente |
| **20** | 9 | 3 | 7 | 2 | 16 | 5 | 6 | 5 | 25 | 7 | regular | bueno |
| **21** | 10 | 5 | 6 | 1 | 18 | 6 | 9 | 3 | 20 | 6 | malo | bueno |
| **22** | 9 | 3 | 9 | 2 | 25 | 5 | 8 | 5 | 20 | 9 | malo | bueno |
| **23** | 8 | 5 | 8 | 3 | 20 | 6 | 8 | 5 | 20 | 8 | regular | bueno |
| **24** | 7 | 5 | 7 | 4 | 16 | 5 | 8 | 5 | 20 | 5 | malo | bueno |
| **25** | 6 | 5 | 8 | 5 | 15 | 4 | 9 | 5 | 25 | 5 | bueno | excelente |
| **26** | 9 | 5 | 9 | 3 | 18 | 5 | 10 | 5 | 22 | 3 | bueno | excelente |
| **27** | 8 | 5 | 7 | 2 | 20 | 4 | 9 | 4 | 25 | 2 | bueno | excelente |
| **28** | 7 | 5 | 5 | 3 | 15 | 5 | 8 | 3 | 20 | 1 | regular | bueno |
| **29** | 8 | 4 | 9 | 5 | 15 | 4 | 7 | 1 | 20 | 2 | regular | bueno |
| **30** | 9 | 3 | 7 | 4 | 15 | 5 | 8 | 3 | 25 | 5 | regular | bueno |

### 4.3.3 Análisis e Interpretación de Resultados

**A. INDICADOR.** Tiempo para verificar datos del alumno: KPI1

**Tabla. Resultados de Pre –Prueba y Post- Prueba para el KPI1.**

Tabla 21
*Kpi 1*

| | PRE-PRUEBA | POST PRUEBA | | |
|---|---|---|---|---|
| | 8 | 2 | 2 | 2 |
| | 9 | 2 | 2 | 2 |
| | 8 | 3 | 3 | 3 |
| | 7 | 4 | 4 | 4 |
| | 6 | 2 | 2 | 2 |
| | 10 | 4 | 4 | 4 |
| | 6 | 3 | 3 | 3 |
| | 7 | 5 | 5 | 5 |
| | 6 | 3 | 3 | 3 |
| | 5 | 4 | 4 | 4 |
| | 6 | 5 | 5 | 5 |
| | 8 | 4 | 4 | 4 |
| | 7 | 3 | 3 | 3 |
| | 6 | 2 | 2 | 2 |
| | 7 | 3 | 3 | 3 |
| | 5 | 2 | 2 | 2 |
| | 8 | 1 | 1 | 1 |
| | 8 | 3 | 3 | 3 |
| | 8 | 5 | 5 | 5 |
| | 9 | 3 | 3 | 3 |
| | 10 | 5 | 5 | 5 |
| | 9 | 3 | 3 | 3 |
| | 8 | 5 | 5 | 5 |
| | 7 | 5 | 5 | 5 |
| | 6 | 5 | 5 | 5 |
| | 9 | 5 | 5 | 5 |
| | 8 | 5 | 5 | 5 |
| | 7 | 5 | 5 | 5 |
| | 8 | 4 | 4 | 4 |
| | 9 | 3 | 3 | 3 |
| **Promedio** | 7,5 | 3,5 | | |
| **Meta Planteada** | | | 5 | |
| **N° menor promedio** | | 21 | 21 | 30 |
| **% menos al promedio** | | 70 | 70 | 100% |

- El 70 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que su tiempo promedio.
- El 70 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que la meta planteada.
- El 100 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que el tiempo promedio en la Pre-Prueba.

**Con estadística descriptiva**

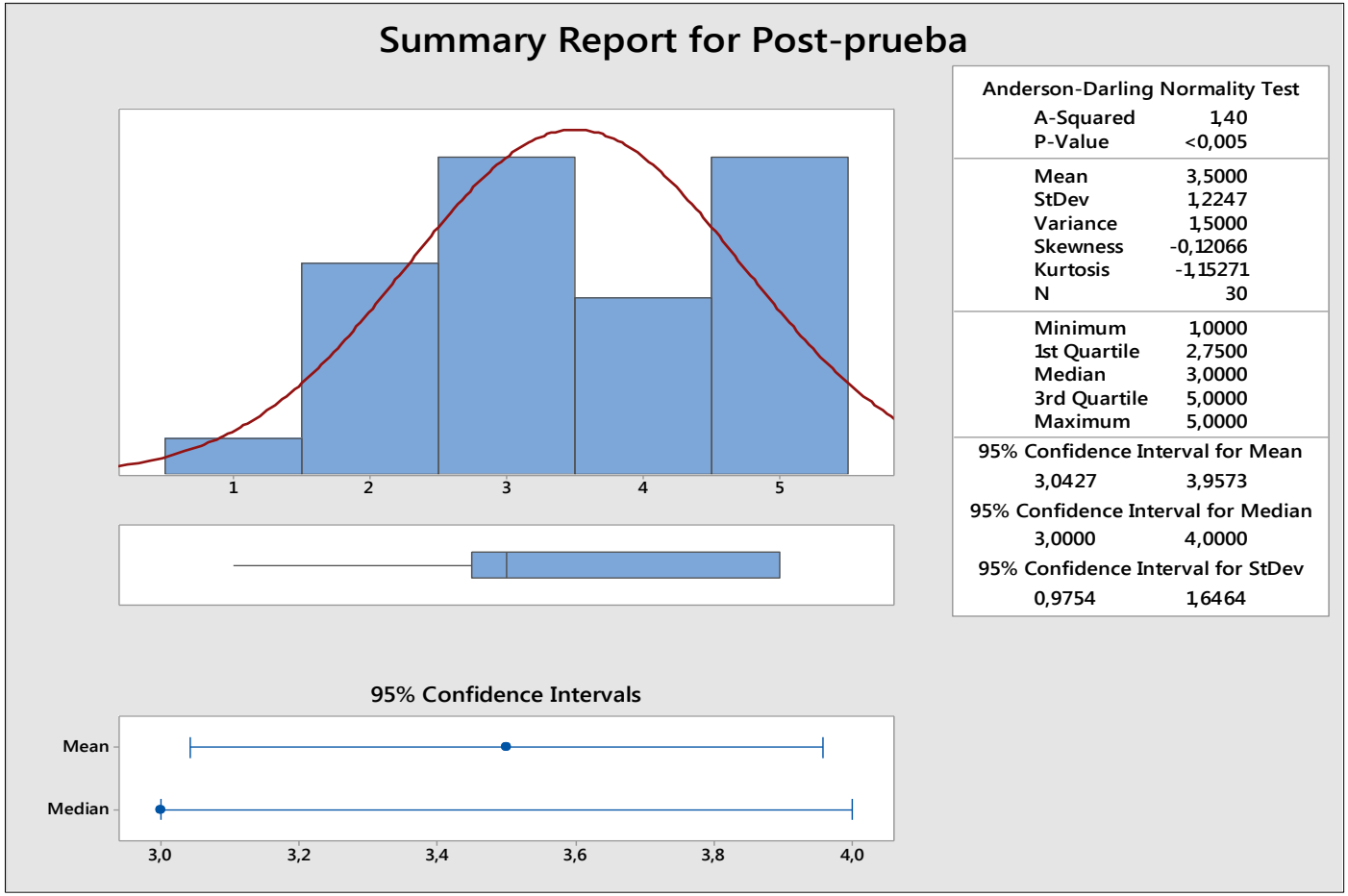


*Figura 69*. Estadística kpi1

- Los datos tienen un comportamiento poco normal debido a que el Valor p (0.005) < α (0.05), pero son valores muy cercanos, lo cual se confirma al observarse que los intervalos de confianza de la Media y la Mediana se traslapan.

- La distancia "promedio" de las observaciones individuales de los Tiempos para registrar las incidencias con respecto a la media es de 1.22 minutos.

- Alrededor del 95% de los Tiempos para registrar las incidencias están dentro de 2 desviaciones estándar de la media, es decir, entre 3.04 y 4.00 minutos.

- La Kurtosis = -1.15 indica que tenemos datos de tiempos con picos muy bajos.

- La Asimetría = -0.12 indica que la mayoría de los Tiempos para registrar las incidencias son bajos.

- El 1er Cuartil (Q1) = 2.750 minutos, indica que el 25% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

- El 3er Cuartil (Q3) = 5.000 minutos, indica que el 75% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

**INDICADOR.** Tiempo para verificar alumno moroso: KPI2

Tabla 22
*Kpi 2*

| | PRE-PRUEBA | POST-PRUEBA | | |
|---|---|---|---|---|
| | 8 | 2 | 2 | 2 |
| | 9 | 3 | 3 | 3 |
| | 8 | 4 | 4 | 4 |
| | 7 | 2 | 2 | 2 |
| | 6 | 1 | 1 | 1 |
| | 10 | 2 | 2 | 2 |
| | 6 | 3 | 3 | 3 |
| | 7 | 4 | 4 | 4 |
| | 6 | 5 | 5 | 5 |
| | 5 | 3 | 3 | 3 |
| | 7 | 3 | 3 | 3 |
| | 5 | 2 | 2 | 2 |
| | 8 | 3 | 3 | 3 |
| | 8 | 4 | 4 | 4 |
| | 8 | 5 | 5 | 5 |
| | 9 | 2 | 2 | 2 |
| | 10 | 3 | 3 | 3 |
| | 9 | 4 | 4 | 4 |
| | 8 | 5 | 5 | 5 |
| | 7 | 5 | 5 | 5 |
| | 6 | 3 | 3 | 3 |
| | 9 | 2 | 2 | 2 |
| | 8 | 1 | 1 | 1 |
| | 7 | 2 | 2 | 2 |
| | 8 | 3 | 3 | 3 |
| | 9 | 4 | 4 | 4 |
| | 7 | 5 | 5 | 5 |
| | 5 | 3 | 3 | 3 |
| | 9 | 2 | 2 | 2 |
| | 7 | 3 | 3 | 3 |
| **Promedio** | 7,43 | | 3,1 | |
| **Meta Planteada** | | | 5 | |
| **N° menor promedio** | | 23 | 23 | 30 |
| **% menos al promedio** | | 77% | 77% | 100% |

- El 77 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que su tiempo promedio.
- El 77 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que la meta planteada.
- El 100 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que el tiempo promedio en la Pre-Prueba

**Con estadística descriptiva**

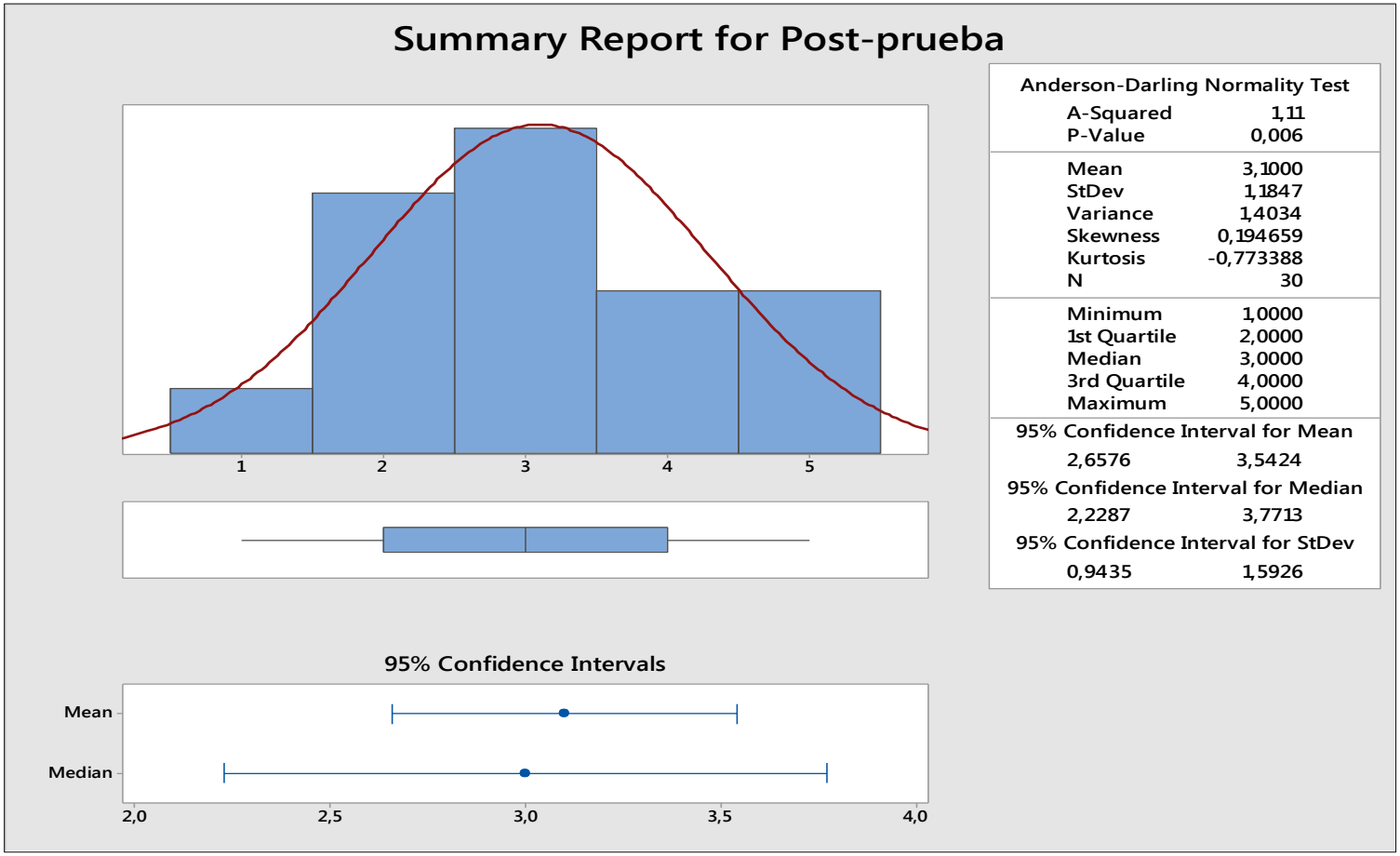


*Figura 70.* Estadística Kpi2

- Los datos tienen un comportamiento poco normal debido a que el Valor p (0.005) < α (0.05), pero son valores muy cercanos, lo cual se confirma al observarse que los intervalos de confianza de la Media y la Mediana se traslapan.

- La distancia "promedio" de las observaciones individuales de los Tiempos para registrar las incidencias con respecto a la media es de 1.18 minutos.

- Alrededor del 95% de los Tiempos para registrar las incidencias están dentro de 2 desviaciones estándar de la media, es decir, entre 2.65 y 3.54 minutos.

- La Kurtosis = -0.77 indica que tenemos datos de tiempos con picos muy bajos.

- La Asimetría = -0.19 indica que la mayoría de los Tiempos para registrar las incidencias son bajos.

- El 1er Cuartil (Q1) = 2.000 minutos, indica que el 25% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

- El 3er Cuartil (Q3) = 4.000 minutos, indica que el 75% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

**INDICADOR.** Índice de Riesgo: KPI3

Tabla 23
*Kpi 3*

| PRE-PRUEBA | POST-PRUEBA | | |
|---|---|---|---|
| 15 | 5 | 5 | 5 |
| 18 | 6 | 6 | 6 |
| 15 | 5 | 5 | 5 |
| 15 | 4 | 4 | 4 |
| 17 | 5 | 5 | 5 |
| 18 | 7 | 7 | 7 |
| 20 | 8 | 8 | 8 |
| 20 | 6 | 6 | 6 |
| 15 | 5 | 5 | 5 |
| 14 | 4 | 4 | 4 |
| 16 | 3 | 3 | 3 |
| 18 | 5 | 5 | 5 |
| 19 | 3 | 3 | 3 |
| 20 | 5 | 5 | 5 |

| | | | | | |
|---|---|---|---|---|---|
| | 25 | 6 | | 6 | 6 |
| | 25 | 5 | | 5 | 5 |
| | 16 | 6 | | 6 | 6 |
| | 18 | 5 | | 5 | 5 |
| | 16 | 6 | | 6 | 6 |
| | 16 | 5 | | 5 | 5 |
| | 18 | 6 | | 6 | 6 |
| | 25 | 5 | | 5 | 5 |
| | 20 | 4 | | 4 | 4 |
| | 16 | 5 | | 5 | 5 |
| | 15 | 4 | | 4 | 4 |
| | 18 | 5 | | 5 | 5 |
| | 20 | 4 | | 4 | 4 |
| | 15 | 5 | | 5 | 5 |
| | 15 | 4 | | 4 | 4 |
| | 15 | 5 | | 5 | 5 |
| **Promedio** | 17,17 | | 5,1 | | |
| **Meta Planteada** | | | | 6 | |
| **N° menor promedio** | | 21 | | 21 | 30 |
| **% menos al promedio** | | 70% | | 70% | 70% |

- 70 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que su tiempo promedio.
- El 70 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que la meta planteada.
- El 100 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que el tiempo promedio en la Pre-Prueba.

**Con estadística descriptiva**

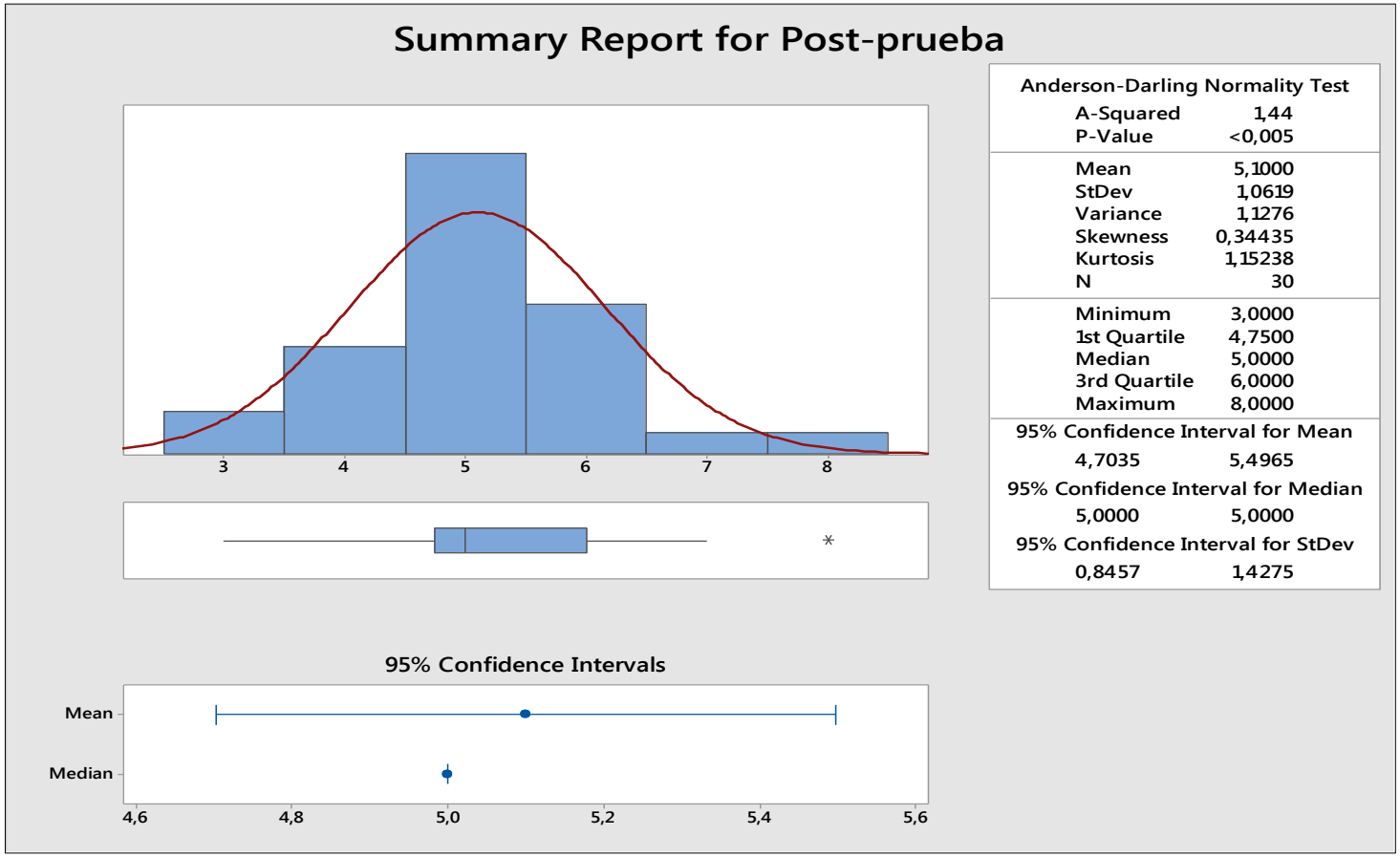


*Figura 7170*. Estadística Kpi3

- Los datos tienen un comportamiento poco normal debido a que el Valor p (0.005) < α (0.05), pero son valores muy cercanos, lo cual se confirma al observarse que los intervalos de confianza de la Media y la Mediana se traslapan.

- La distancia "promedio" de las observaciones individuales de los Tiempos para registrar las incidencias con respecto a la media es de 1.06 minutos.

- Alrededor del 95% de los Tiempos para registrar las incidencias están dentro de 2 desviaciones estándar de la media, es decir, entre 4.7 y 5.49 minutos.

- La Kurtosis = 1.15 indica que tenemos datos de tiempos con picos muy bajos.

- La Asimetría = 0.344 indica que la mayoría de los Tiempos para registrar las incidencias son bajos.

- El 1er Cuartil (Q1) = 4.750 minutos, indica que el 25% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

- El 3er Cuartil (Q3) = 6.000 minutos, indica que el 75% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

**INDICADOR.** Tiempo para determinar el riesgo de morosidad de un estudiante: KPI4

Tabla 24
*Estadística Kpi4*

| PRE-PRUEBA | POST-PRUEBA | | |
|---|---|---|---|
| 8 | 4 | 4 | 4 |
| 9 | 5 | 5 | 5 |
| 8 | 3 | 3 | 3 |
| 7 | 2 | 2 | 2 |
| 6 | 3 | 3 | 3 |
| 10 | 5 | 5 | 5 |
| 6 | 2 | 2 | 2 |
| 7 | 2 | 2 | 2 |
| 6 | 3 | 3 | 3 |
| 5 | 4 | 4 | 4 |
| 5 | 2 | 2 | 2 |
| 8 | 4 | 4 | 4 |
| 8 | 3 | 3 | 3 |
| 8 | 5 | 5 | 5 |
| 9 | 3 | 3 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| | 10 | 4 | | 4 | 4 |
| | 9 | 4 | | 4 | 4 |
| | 8 | 5 | | 5 | 5 |
| | 7 | 4 | | 4 | 4 |
| | 6 | 5 | | 5 | 5 |
| | 9 | 3 | | 3 | 3 |
| | 8 | 5 | | 5 | 5 |
| | 8 | 5 | | 5 | 5 |
| | 8 | 5 | | 5 | 5 |
| | 9 | 5 | | 5 | 5 |
| | 10 | 5 | | 5 | 5 |
| | 9 | 4 | | 4 | 4 |
| | 8 | 3 | | 3 | 3 |
| | 7 | 1 | | 1 | 1 |
| | 8 | 3 | | 3 | 3 |
| **Promedio** | 7,80 | | 3,7 | | |
| **Meta Planteada** | | | | 5 | |
| **N° menor promedio** | | 20 | | 20 | 20 |
| **% menos al promedio** | | 67% | | 67% | 100% |

- El 67 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que su tiempo promedio.
- El 67 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que la meta planteada.
- El 100 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que el tiempo promedio en la Pre-Prueba.

**Con estadística descriptiva**

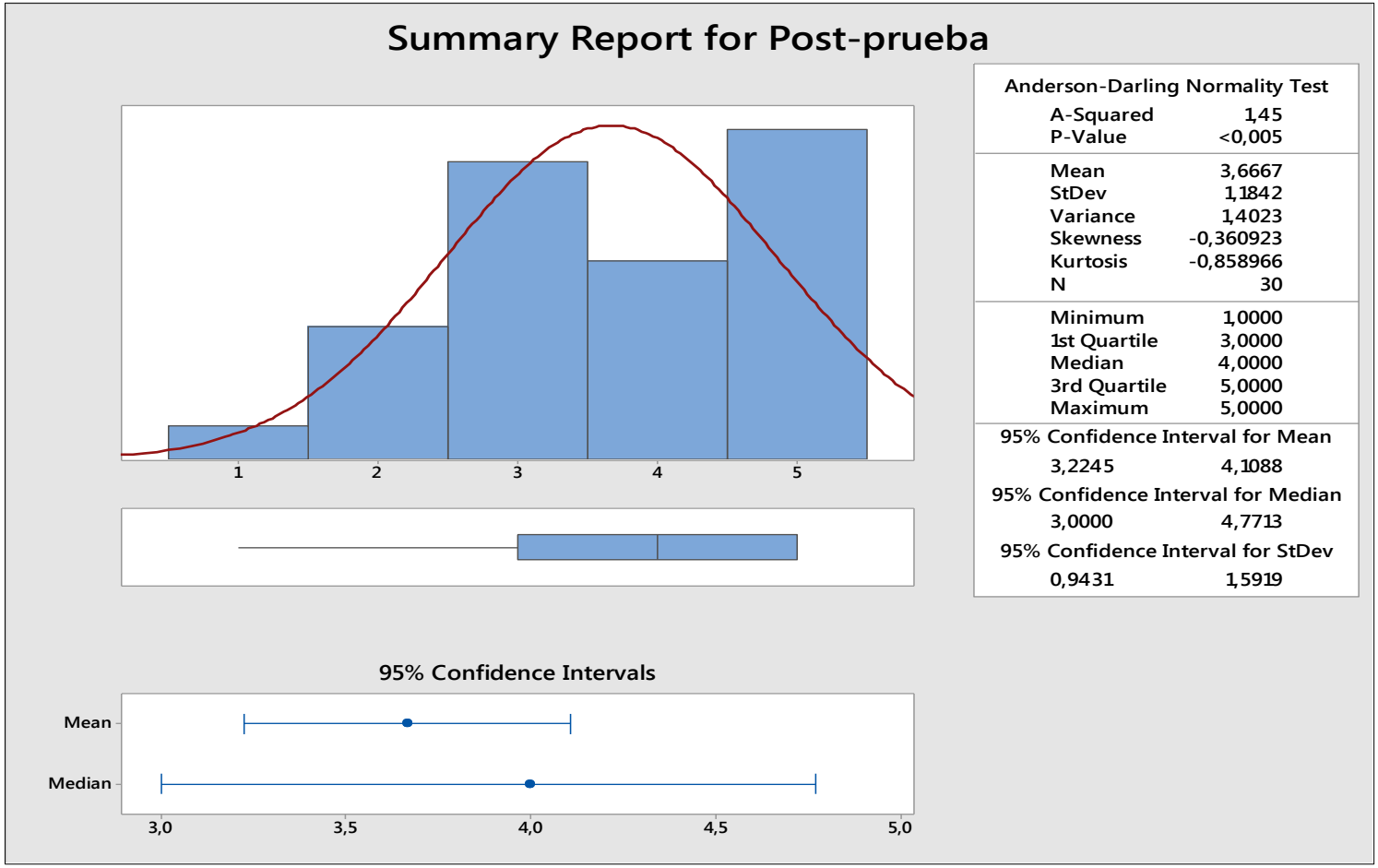


*Figura 72.* Estadística Kpi4

- Los datos tienen un comportamiento poco normal debido a que el Valor p (0.005) < α (0.05), pero son valores muy cercanos, lo cual se confirma al observarse que los intervalos de confianza de la Media y la Mediana se traslapan.

- La distancia "promedio" de las observaciones individuales de los Tiempos para registrar las incidencias con respecto a la media es de 1.18 minutos.

- Alrededor del 95% de los Tiempos para registrar las incidencias están dentro de 2 desviaciones estándar de la media, es decir, entre 3.22 y 4.10 minutos.

- La Kurtosis = -0.85 indica que tenemos datos de tiempos con picos muy bajos.

- La Asimetría = -0.360 indica que la mayoría de los Tiempos para registrar las incidencias son bajos.

- El 1er Cuartil (Q1) = 3.000 minutos, indica que el 25% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

- El 3er Cuartil (Q3) = 5.000 minutos, indica que el 75% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

**INDICADOR.** Tiempo para predecir que alumnos incurrirán en morosidad: KPI5

Tabla 25
*Estadística Kpi5*

| PRE-PRUEBA | POST-PRUEBA | | |
|---|---|---|---|
| 20 | 3 | 3 | 3 |
| 25 | 2 | 2 | 2 |
| 30 | 3 | 3 | 3 |
| 26 | 5 | 5 | 5 |
| 25 | 2 | 2 | 2 |
| 20 | 2 | 2 | 2 |
| 30 | 3 | 3 | 3 |
| 28 | 4 | 4 | 4 |
| 25 | 2 | 2 | 2 |
| 20 | 4 | 4 | 4 |
| 20 | 4 | 4 | 4 |
| 20 | 2 | 2 | 2 |
| 25 | 1 | 1 | 1 |
| 25 | 2 | 2 | 2 |
| 20 | 3 | 3 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| | 30 | 4 | | 4 | 4 |
| | 25 | 5 | | 5 | 5 |
| | 20 | 3 | | 3 | 3 |
| | 20 | 8 | | 8 | 8 |
| | 25 | 7 | | 7 | 7 |
| | 20 | 6 | | 6 | 6 |
| | 20 | 9 | | 9 | 9 |
| | 20 | 8 | | 8 | 8 |
| | 20 | 5 | | 5 | 5 |
| | 25 | 5 | | 5 | 5 |
| | 22 | 3 | | 3 | 3 |
| | 25 | 2 | | 2 | 2 |
| | 20 | 1 | | 1 | 1 |
| | 20 | 2 | | 2 | 2 |
| | 25 | 5 | | 5 | 5 |
| **Promedio** | 23,2 | | 3,83 | | |
| **Meta Planteada** | | | | 6 | |
| **N° menor promedio** | | 25 | | 25 | 30 |
| **% menos al promedio** | | 83% | | 83% | 100% |

- El 83 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que su tiempo promedio.
- El 83 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que la meta planteada.
- El 100 % de los Tiempos para registrar las incidencias en la Post-Prueba fueron menores que el tiempo promedio en la Pre-Prueba.

**Con estadística descriptiva**

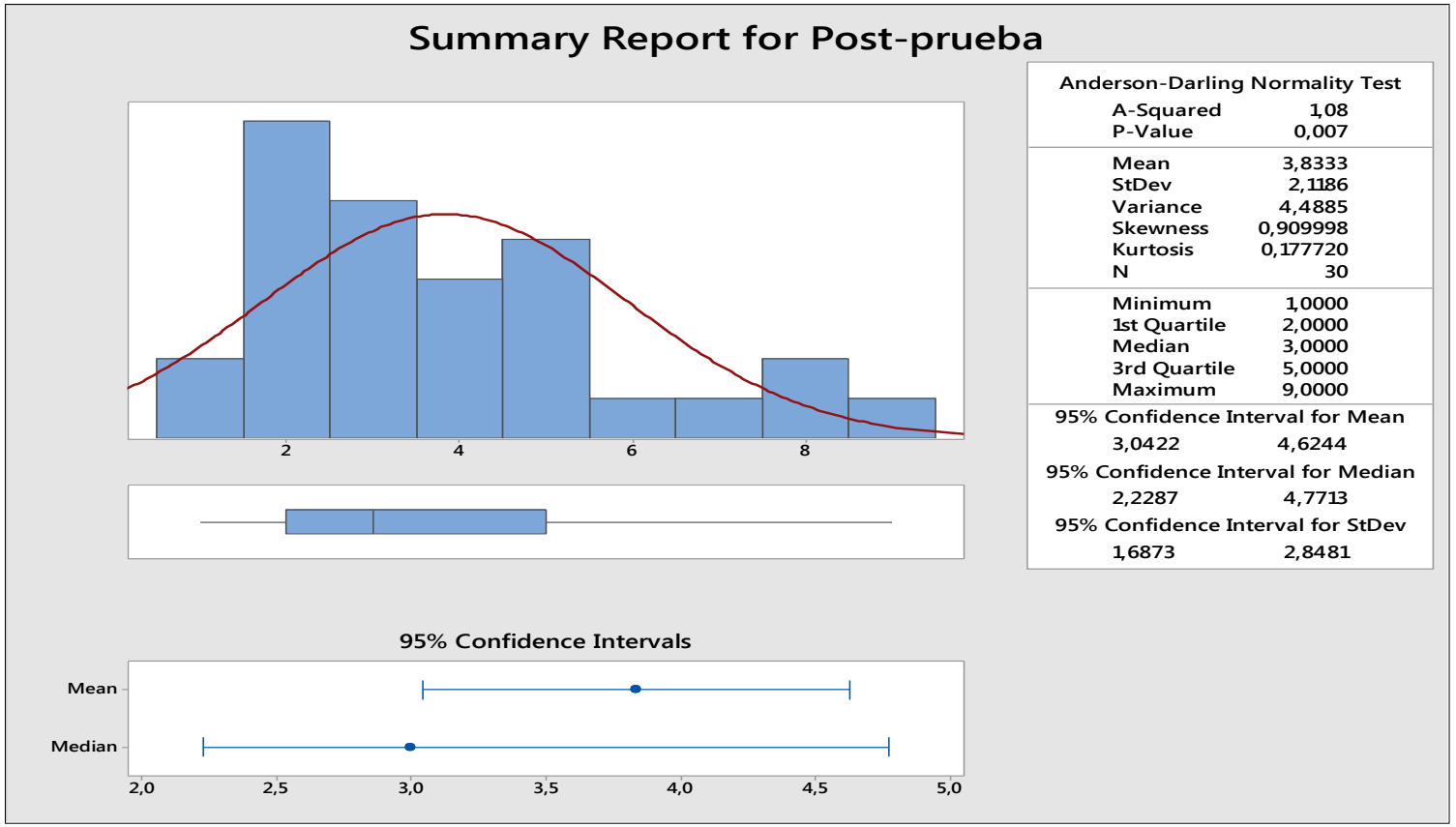


*Figura 73.* Estadística Kpi5

- Los datos tienen un comportamiento poco normal debido a que el Valor p (0.005) < α (0.05), pero son valores muy cercanos, lo cual se confirma al observarse que los intervalos de confianza de la Media y la Mediana se traslapan.

- La distancia "promedio" de las observaciones individuales de los Tiempos para registrar las incidencias con respecto a la media es de 2.11 minutos.

- Alrededor del 95% de los Tiempos para registrar las incidencias están dentro de 2 desviaciones estándar de la media, es decir, entre 3.04 y 4.62 minutos.

- La Kurtosis = 0.17 indica que tenemos datos de tiempos con picos muy bajos.

- La Asimetría = 0.909 indica que la mayoría de los Tiempos para registrar las incidencias son bajos.

- El 1er Cuartil (Q1) = 2.000 minutos, indica que el 25% de los Tiempos para registrar las incidencias es menor que o igual a este valor.

- El 3er Cuartil (Q3) = 5.000 minutos, indica que el 75% de los Tiempos para registrar las incidencias es menor que o igual a este valor

**INDICADOR.** Nivel de Satisfacción: KPI6

### Valores de la Pre-Prueba

Tabla 27
*Estadística Kpi6*

| | **Post-prueba** |
|---|---|
| **1** | bueno |
| **2** | regular |
| **3** | regular |
| **4** | bueno |
| **5** | malo |
| **6** | malo |

Tabla 26
*Estadística Kpi6*

| **Estado** | **Frecuencia** |
|---|---|
| Malo | 7 |
| Regular | 12 |
| Bueno | 11 |
| Excelente | 0 |
| TOTAL | 30 |

| | |
|---|---|
| **7** | bueno |
| **8** | regular |
| **9** | regular |
| **10** | bueno |
| **11** | regular |
| **12** | malo |
| **13** | regular |
| **14** | malo |
| **15** | regular |
| **16** | bueno |
| **17** | bueno |
| **18** | bueno |
| **19** | bueno |
| **20** | malo |
| **21** | malo |
| **22** | regular |
| **23** | malo |
| **24** | bueno |
| **25** | bueno |
| **26** | bueno |
| **27** | regular |
| **28** | regular |
| **29** | regular |
| **30** | regular |

Tabla 28
*Estadística Kpi6*

| **Estado** | **Frecuencia** |
|---|---|
| Malo | 19 |
| Bueno | 11 |

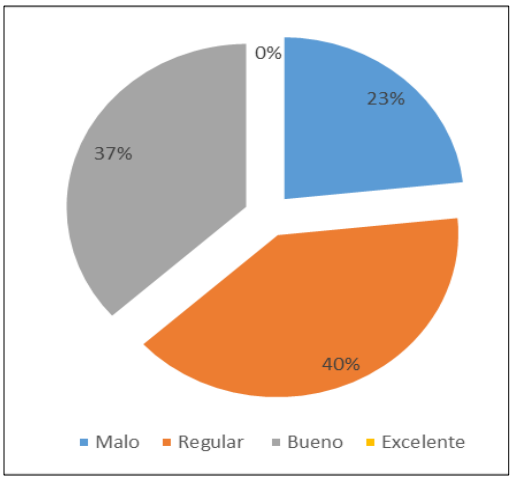


*Figura 714*. Estadística valor kpi6

- El 23.0 % de las veces el Nivel de Satisfacción fue catalogado como Malo por los usuarios atendidos.
- El 37.0 % de las veces el Nivel de Satisfacción fue catalogado como Bueno por los usuarios atendidos.
- Se determina que sólo el 37.0 % de las veces el Nivel de Satisfacción es Buena.
- Se determina que el 63.0 % de las veces el Nivel de Satisfacción es Mala

## Valores de la Post-Prueba:

Tabla 29
*Estadística Kpi7*

| | Post-prueba |
|---|---|
| **1** | exelente |
| **2** | bueno |
| **3** | bueno |
| **4** | exelente |
| **5** | bueno |
| **6** | regular |
| **7** | exelente |
| **8** | bueno |
| **9** | bueno |
| **10** | exelente |
| **11** | bueno |
| **12** | bueno |
| **13** | bueno |
| **14** | bueno |
| **15** | bueno |
| **16** | exelente |
| **17** | exelente |
| **18** | exelente |
| **19** | exelente |
| **20** | bueno |
| **21** | bueno |
| **22** | bueno |
| **23** | bueno |
| **24** | bueno |
| **25** | exelente |
| **26** | exelente |
| **27** | exelente |
| **28** | bueno |
| **29** | bueno |
| **30** | bueno |

Tabla 30
Estadística Kpi7

| Estado | Frecuencia |
|---|---|
| Malo | 0 |
| Regular | 1 |
| Bueno | 18 |
| Excelente | 11 |
| Total | 30 |

Tabla 31
*Estadística Kpi7*

| Estado | Frecuencia |
|---|---|
| Bueno | 29 |
| Malo | 1 |

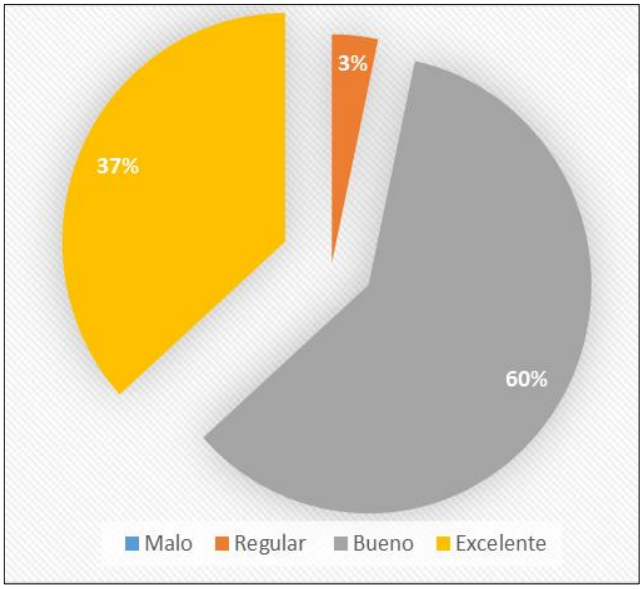


*Figura 75*. Estadística valor post kpi7

- El 3.0 % de las veces el Nivel de Satisfacción fue catalogada como Regular por los usuarios atendidos.
- El 37.0 % de las veces el Nivel de Satisfacción fue catalogada como Excelente por los usuarios atendidos.
- Se determina ahora que el 60.0 % de las veces el Nivel de Satisfacción es Buena.
- Se determina ahora que sólo el 0.0 % de las veces el Nivel de Satisfacción es Mala.

## 4.4 PRUEBA DE HIPÓTESIS

A continuación, se presentan las medias de los KPIs para la Pre-Prueba y Post-Prueba: Resultados numéricos

Tabla 32
*KPI*

| Indicador | Pre -Prueba (Media: $\boldsymbol{X}1$) | Pros -Prueba (Media: $\boldsymbol{X}2$) |
|---|---|---|
| **Tiempo para identificar alumno moroso** | 5 minutos/alumno | 2 minutos/alumno |
| **Tiempo para determinar cualidades del alumno moroso** | 60 minutos/alumno | 5 minutos/alumno |
| **Tiempo para verificar alumno moroso** | 25 minutos/alumno | 3 minutos/alumno |
| **Tiempo para predecir al alumno moroso** | 15 minutos/solicitud | 2 minutos/solicitud |
| **Tiempo para verificar porcentaje de alumnos morosos** | 30 minutos/ | 5 minutos/ |
| **Tiempo para generar exactitud de reportes de los alumnos morosos** | 20 minutos/ | 4 minutos/ |
| **Nivel de Satisfacción** | Regular | Exelente |

### 4.4.1 Prueba para el indicador: Tiempo para verificar datos del alumno- KPI1

Se valida el impacto que tiene al realizar un modelo predictivo en el Tiempo para verificar datos del alumno, llevado a cabo en la muestra. Se realizó una medición

antes de realizar el modelo predictivo (Post-Prueba O2) y otra después de realizar un modelo predictivo (Post-Prueba O1).

Hipótesis Específica Hi1: Si se realizar el modelo predictivo para el proceso de refinanciamiento de los alumnos en la Universidad Autónoma del Perú, se disminuirá el tiempo para verificar datos del alumno (Post-Prueba O1) con respecto a la muestra a la cual se realiza (Post-Prueba O2).

Tabla 33
*Prueba Kpi1*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 8 | 9 | 8 | 7 | 6 | 10 | 6 |
| | 7 | 6 | 5 | 7 | 6 | 7 | 5 |
| **Pre - Prueba** | 8 | 8 | 8 | 9 | 10 | 9 | 8 |
| | 7 | 6 | 9 | 8 | 7 | 8 | 9 |
| | 7 | 5 | | | | | |

Tabla 34
*Prueba Kpi1*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 2 | 1 | 2 | 3 |
| | 4 | 5 | 3 | 3 | 2 | 3 | 4 |
| **Pre - Prueba** | 5 | 2 | 3 | 4 | 5 | 5 | 3 |
| | 2 | 1 | 2 | 3 | 4 | 5 | 3 |
| | 2 | 3 | | | | | |

**Hi1:** El uso de un modelo predictivo disminuye el tiempo para verificar datos del alumno **(Post-Prueba)** con respecto a la muestra a la que no se aplicó **(Pre-Prueba).**

**Solución:**

**a) Planteamiento de la Hipótesis**

**μ1=** Media del Tiempo para registrar las incidencias Pre-Prueba.

**μ2=** Media del Tiempo para registrar las incidencias Post-Prueba

**b) Criterio de Decisión:**

$$H_o: \mu_1 \leq \mu_2$$
$$H_a: \mu_1 > \mu_2$$

**c) Cálculo: Prueba t para prueba de medias de las dos muestras:**

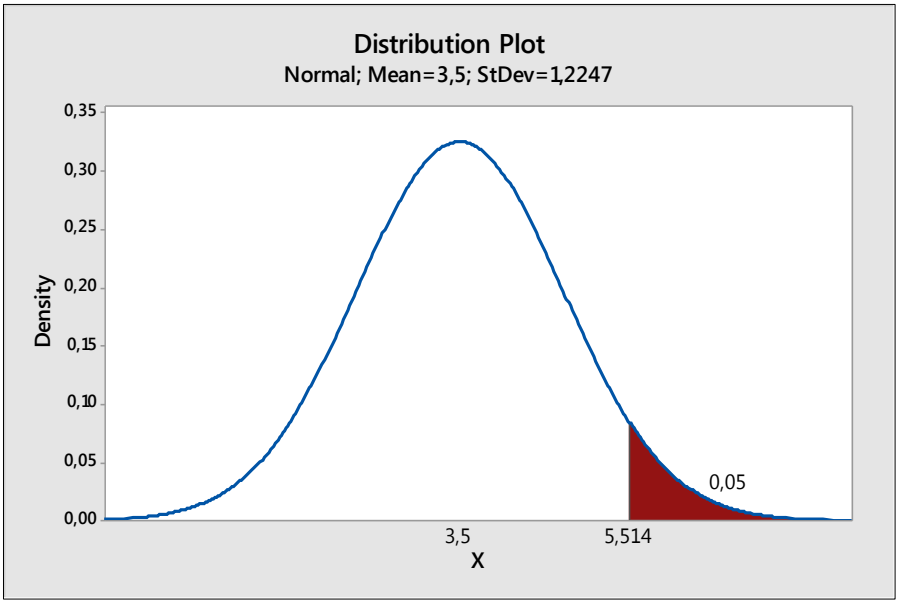


*Figura 76.* Estadística valor post kpi7

Tabla 35
*Tiempo para verificar datos del alumno del KPI1.*

| | **Variable 1** | **Variable 2** |
|---|---|---|
| Media | 7.5 | 3.5 |
| Varianza | 1.84482759 | 1.5 |
| Observaciones | 30 | 30 |
| Coeficiente de correlacion de Pearson | 0.0725515220 | |
| Diferencia hipotética de las medias | 0 | |
| Grados de libertad | 29 | |
| Estadistico t | 12.4365054 | |
| P(T<=T) una cola | 1.89245E-13 | |
| Valor critico de t (una cola) | 1.699127027 | |
| P(T<=T) dos cola | 3.7849E-13 | |
| Valor critico de t (dos colas) | 2.045229642 | |

**a) Decisión Estadística**

Puesto que el valor-p=0.000 < α=0.05, los resultados proporcionan suficiente evidencia para rechazar la hipótesis nula (Ho), y la hipótesis alterna (Ha) es cierta. La prueba resultó ser significativa.

### 4.4.2 Prueba para el indicador: Tiempo para registrar el monto a pagar KPI2

Se valida el impacto que tiene al realizar un modelo predictivo en el Tiempo para registrar el monto a pagar, llevado a cabo en la muestra. Se realizó una medición antes de realizar el modelo predictivo (Post-Prueba O2) y otra después de realizar un modelo predictivo (Post-Prueba O1).

Hipótesis Específica Hi2: Si se realizar el modelo predictivo para el proceso de refinanciamiento de los alumnos en la Universidad Autónoma del Perú, se disminuirá el tiempo para registrar el monto a pagar (Post-Prueba O1) con respecto a la muestra a la cual se realiza (Post-Prueba O2).

**Hi2:** El uso de un modelo predictivo disminuye el tiempo para registrar el monto a pagar **(Post-Prueba)** con respecto a la muestra a la que no se aplicó **(Pre-Prueba).**

Tabla 36

*Prueba Kpi2*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 8 | 9 | 8 | 7 | 6 | 10 | 6 |
| | 7 | 6 | 5 | 7 | 6 | 7 | 5 |
| **Pre - Prueba** | 8 | 8 | 8 | 9 | 10 | 9 | 8 |
| | 7 | 6 | 9 | 8 | 7 | 8 | 9 |
| | 7 | 5 | | | | | |

Tabla 37

*Prueba Kpi2*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 2 | 1 | 2 | 3 |
| | 4 | 5 | 3 | 3 | 2 | 3 | 4 |
| **Pre - Prueba** | 5 | 2 | 3 | 4 | 5 | 5 | 3 |
| | 2 | 1 | 2 | 3 | 4 | 5 | 3 |
| | 2 | 3 | | | | | |

**Solución:**

**a) Planteamiento de la Hipótesis**

**μ1=** Media del Tiempo para registrar las incidencias Pre-Prueba.

**μ2=** Media del Tiempo para registrar las incidencias Post-Prueba

$$H_o: \mu_1 \leq \mu_2$$

$$H_a: \mu_1 > \mu_2$$

## b) Criterio de Decisión:

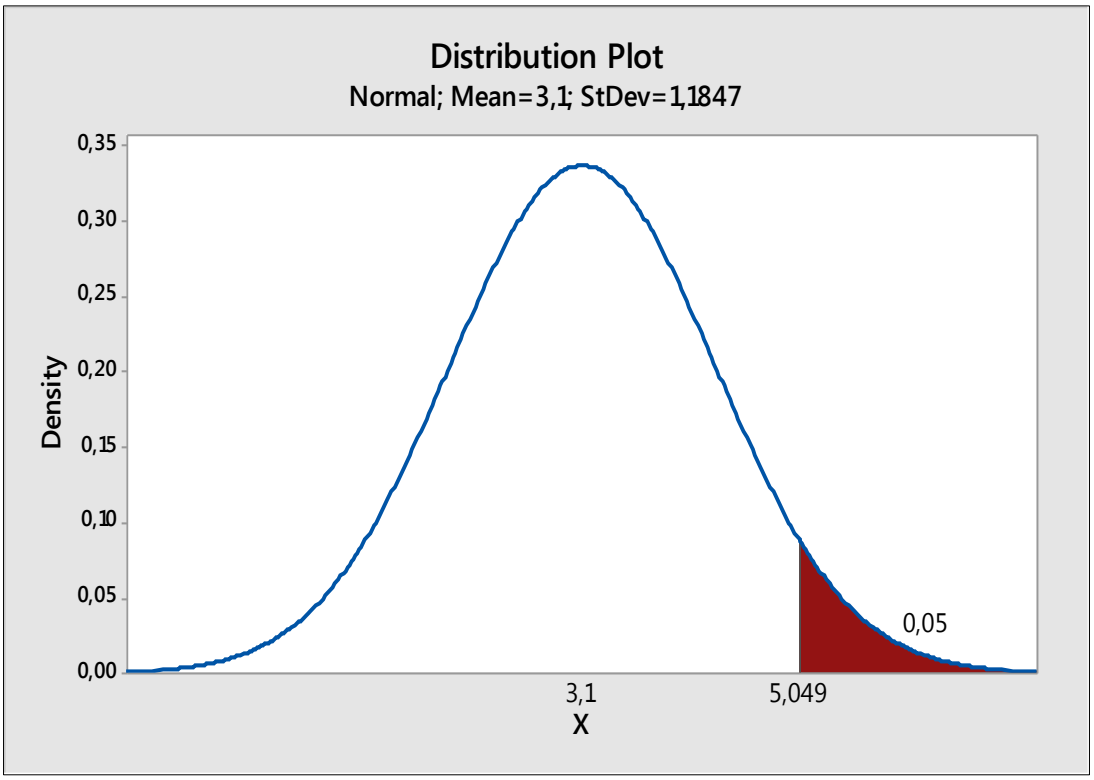


*Figura 77*. Distribución 2

## c) Cálculo: Prueba t para prueba de medias de las dos muestras

Tabla 38
*Tiempo para verificar datos del alumno del KPI2.*

| | *Variable 1* | *Variable 2* |
|---|---|---|
| Media | 7,43333333 | 3,1 |
| Varianza | 1,97816092 | 1,40344828 |
| Observaciones | 30 | 30 |
| Coeficiente de correlación de Pearson | 0,20074499 | |
| Diferencia hipotética de las medias | 0 | |
| Grados de libertad | 29 | |
| Estadístico t | 14,4107422 | |
| P(T<=t) una cola | 4,7003E-15 | |
| Valor crítico de t (una cola) | 1,69912703 | |
| P(T<=t) dos colas | 9,4006E-15 | |
| Valor crítico de t (dos colas) | 2,04522964 | |

**d) Decisión Estadística**

Puesto que el valor-p=0.000 < α=0.05, los resultados proporcionan suficiente evidencia para rechazar la hipótesis nula (Ho), y la hipótesis alterna (Ha) es cierta. La prueba resultó ser significativa.

Tabla 39

*Prueba Kpi3*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 15 | 18 | 15 | 15 | 17 | 18 | 20 |
| | 20 | 15 | 14 | 16 | 18 | 19 | 20 |
| **Pre - Prueba** | 25 | 25 | 16 | 18 | 16 | 16 | 18 |
| | 25 | 20 | 16 | 15 | 18 | 20 | 15 |
| | 15 | 15 | | | | | |

Tabla 40

*Prueba Kpi3*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 5 | 6 | 5 | 4 | 5 | 7 | 8 |
| | 6 | 5 | 4 | 3 | 5 | 3 | 5 |
| **Post - Prueba** | 6 | 5 | 6 | 5 | 6 | 5 | 6 |
| | 5 | 6 | 5 | 4 | 5 | 4 | 5 |
| | 4 | 5 | | | | | |

### 4.4.3 Prueba para el indicador: Tiempo para verificar alumno moroso KPI3

Se valida el impacto que tiene al realizar un modelo predictivo en el Tiempo para verificar alumno moroso, llevado a cabo en la muestra. Se realizó una medición antes de realizar el modelo predictivo (Post-Prueba O2) y otra después de realizar un modelo predictivo (Post-Prueba O1).

Hipótesis Específica Hi3: Si se realizar el modelo predictivo para el proceso de refinanciamiento de los alumnos en la Universidad Autónoma del Perú, se disminuirá el tiempo para verificar alumno moroso (Post-Prueba O1) con respecto a la muestra a la cual se realiza (Post-Prueba O2).

**Hi3:** El uso de un modelo predictivo disminuye el tiempo para para verificar alumno moroso **(Post-Prueba)** con respecto a la muestra a la que no se aplicó **(Pre-Prueba).**

**Solución:**

**a) Planteamiento de la Hipótesis**

**µ1=** Media del Tiempo para registrar las incidencias Pre-Prueba.

**µ2=** Media del Tiempo para registrar las incidencias Post-Prueba

$H_o: \mu_1 \leq \mu_2$

$H_a: \mu_1 > \mu_2$

**b) Criterio de Decisión**

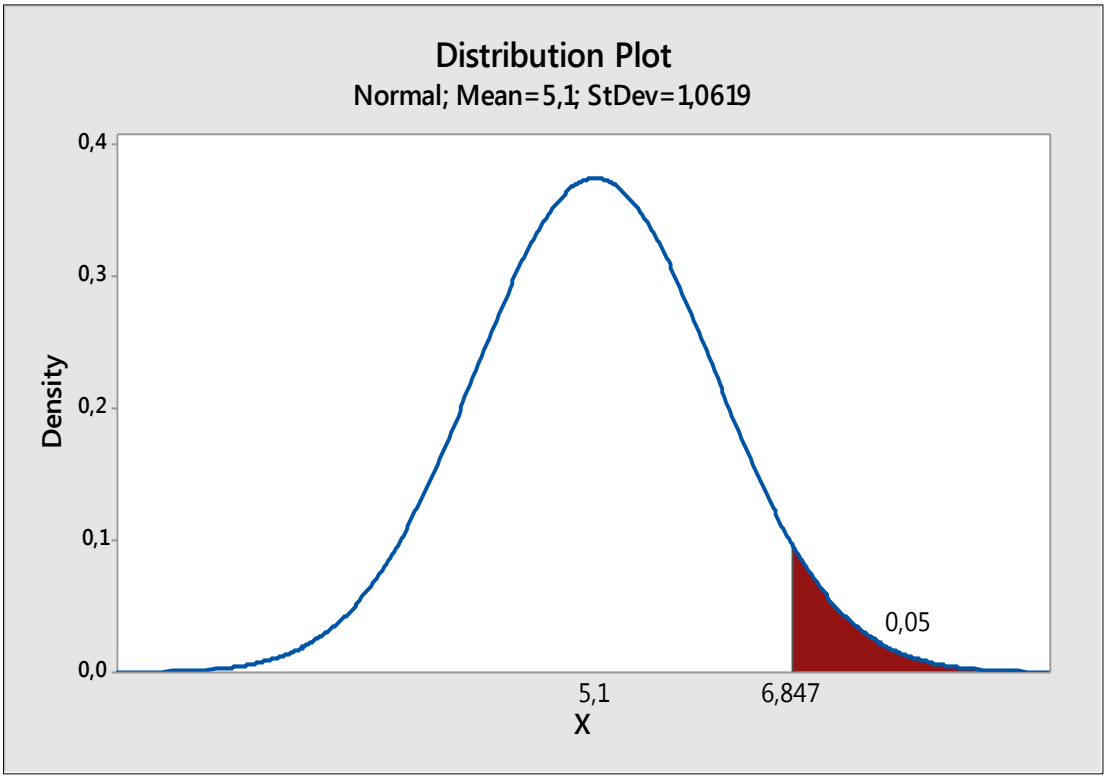


*Figura 78.* Distribución 3

**c) Cálculo: Prueba t para prueba de medias de las dos muestras**

Tabla 41
*Tiempo para verificar datos del alumno del KPI3*

| | *Variable 1* | *Variable 2* |
|---|---|---|
| Media | 17,7666667 | 5,1 |
| Varianza | 9,4954023 | 1,12758621 |
| Observaciones | 30 | 30 |
| Coeficiente de correlación de Pearson | 0,28137253 | |

| | |
|---|---|
| Diferencia hipotética de las medias | 0 |
| Grados de libertad | 29 |
| Estadístico t | 23,4118599 |
| P(T<=t) una cola | 1,1038E-20 |
| Valor crítico de t (una cola) | 1,69912703 |
| P(T<=t) dos colas | 2,2076E-20 |
| Valor crítico de t (dos colas) | 2,04522964 |

**d) Decisión Estadística**

Puesto que el valor-p=0.000 < α=0.05, los resultados proporcionan suficiente evidencia para rechazar la hipótesis nula (Ho), y la hipótesis alterna (Ha) es cierta. La prueba resultó ser significativa.

### 4.4.5 Prueba para el indicador: Tiempo para procesar información KPI4

Se valida el impacto que tiene al realizar un modelo predictivo en el Tiempo para procesar información, llevado a cabo en la muestra. Se realizó una medición antes de realizar el modelo predictivo (Post-Prueba O2) y otra después de realizar un modelo predictivo (Post-Prueba O1).

Hipótesis Específica Hi4: Si se realizar el modelo predictivo para el proceso de refinanciamiento de los alumnos en la Universidad Autónoma del Perú, se disminuirá el para procesar información (Post-Prueba O1) con respecto a la muestra a la cual se realiza (Post-Prueba O2).

Tabla 42

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 8 | 9 | 8 | 7 | 6 | 10 | 6 |
| | 7 | 6 | 5 | 5 | 8 | 8 | 8 |
| **Pre - Prueba** | 9 | 10 | 9 | 8 | 7 | 6 | 9 |
| | 8 | 8 | 8 | 9 | 10 | 9 | 8 |
| | 7 | 8 | | | | | |

*Prueba Kpi4*

Tabla 43
Prueba Kpi4

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 4 | 5 | 3 | 3 | 2 | 3 | 5 |
| | 2 | 2 | 3 | 4 | 2 | 4 | 3 |
| **Pre - Prueba** | 5 | 3 | 4 | 5 | 4 | 5 | 3 |
| | 5 | 5 | 5 | 5 | 5 | 4 | 3 |
| | 1 | 3 | | | | | |

**Hi4:** El uso de un modelo predictivo disminuye el tiempo para para procesar información **(Post-Prueba)** con respecto a la muestra a la que no se aplicó **(Pre-Prueba).**

**Solución:**

**a) Planteamiento de la Hipótesis**

**μ1=** Media del Tiempo para registrar las incidencias Pre-Prueba.

**μ2=** Media del Tiempo para registrar las incidencias Post-Prueba

$$H_o: \mu_1 \leq \mu_2$$

$$H_a: \mu_1 > \mu_2$$

**b) Criterio de Decisión:**

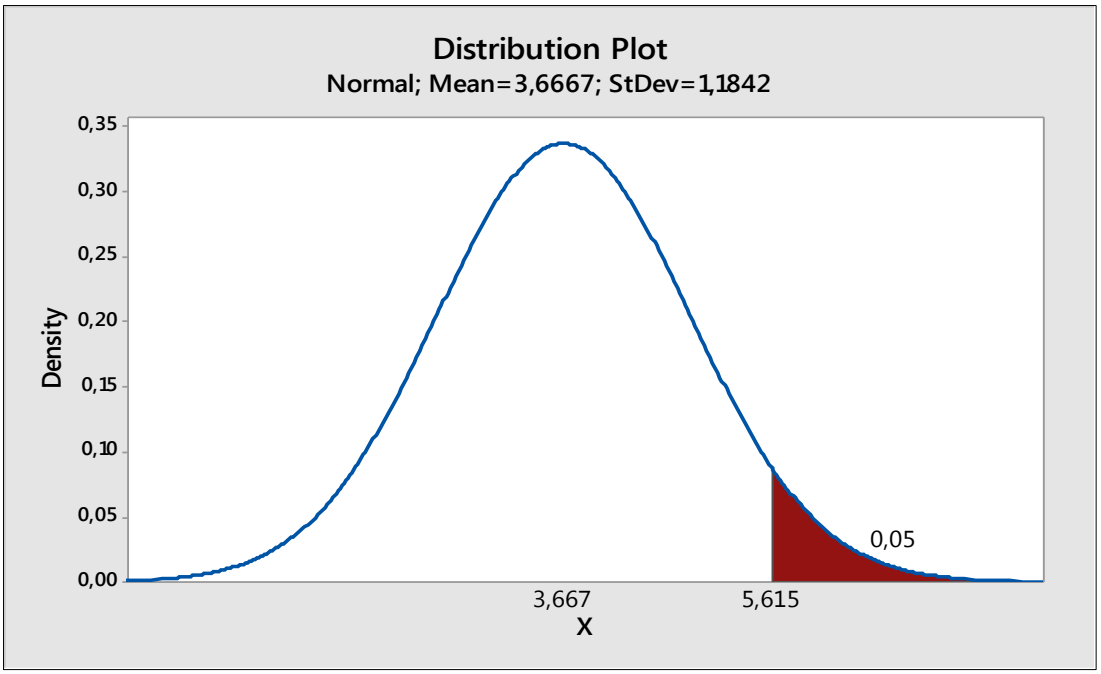


*Figura 7972*. Distribución 4

**c) Cálculo: Prueba t para prueba de medias de las dos muestras**

Tabla 44
*Tiempo para para procesar información del KPI4.*

| | *Variable 1* | *Variable 2* |
|---|---|---|
| Media | 7,8 | 3,66666667 |
| Varianza | 1,88965517 | 1,40229885 |
| Observaciones | 30 | 30 |
| Coeficiente de correlación de Pearson | 0,23301459 | |
| Diferencia hipotética de las medias | 0 | |
| Grados de libertad | 29 | |
| Estadístico t | 14,2237755 | |
| P(T<=t) una cola | 6,5594E-15 | |
| Valor crítico de t (una cola) | 1,69912703 | |
| P(T<=t) dos colas | 1,3119E-14 | |
| Valor crítico de t (dos colas) | 2,04522964 | |

**d) Decisión Estadística**

Puesto que el valor-p=0.000 < α=0.05, los resultados proporcionan suficiente evidencia para rechazar la hipótesis nula (Ho), y la hipótesis alterna (Ha) es cierta. La prueba resultó ser significativa.

### 4.3.5 Prueba para el indicador: Tiempo para procesar información KPI5

Se valida el impacto que tiene al realizar un modelo predictivo en el Tiempo para procesar información, llevado a cabo en la muestra. Se realizó una medición antes de realizar el modelo predictivo (Post-Prueba O2) y otra después de realizar un modelo predictivo (Post-Prueba O1).

Hipótesis Específica Hi5: Si se realizar el modelo predictivo para el proceso de refinanciamiento de los alumnos en la Universidad Autónoma del Perú, se disminuirá el para procesar información (Post-Prueba O1) con respecto a la muestra a la cual se realiza (Post-Prueba O2).

Tabla 45
Estadística Kpi5

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 20 | 25 | 30 | 26 | 25 | 20 | 30 |
| | 28 | 25 | 20 | 20 | 20 | 25 | 25 |
| **Pre - Prueba** | 20 | 30 | 25 | 20 | 20 | 25 | 20 |
| | 20 | 20 | 20 | 20 | 25 | 22 | 25 |
| | 20 | 25 | | | | | |

Tabla 46
*Estadística Kpi5*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 3 | 2 | 3 | 5 | 2 | 2 | 3 |
| | 4 | 2 | 4 | 4 | 2 | 1 | 2 |
| **Pre - Prueba** | 4 | 8 | 3 | 4 | 5 | 3 | 8 |
| | 9 | 4 | 5 | 5 | 3 | 2 | 1 |
| | 2 | 5 | | | | | |

**Hi5:** El uso de un modelo predictivo disminuye el tiempo para procesar información **(Post-Prueba)** con respecto a la muestra a la que no se aplicó **(Pre-Prueba).**

**Solución:**

**a) Planteamiento de la Hipótesis**

**μ1=** Media del Tiempo para registrar las incidencias Pre-Prueba.

**μ2=** Media del Tiempo para registrar las incidencias Post-Prueba

$$H_o: \mu_1 \leq \mu_2$$
$$H_a: \mu_1 > \mu_2$$

## b) Criterio de Decisión:

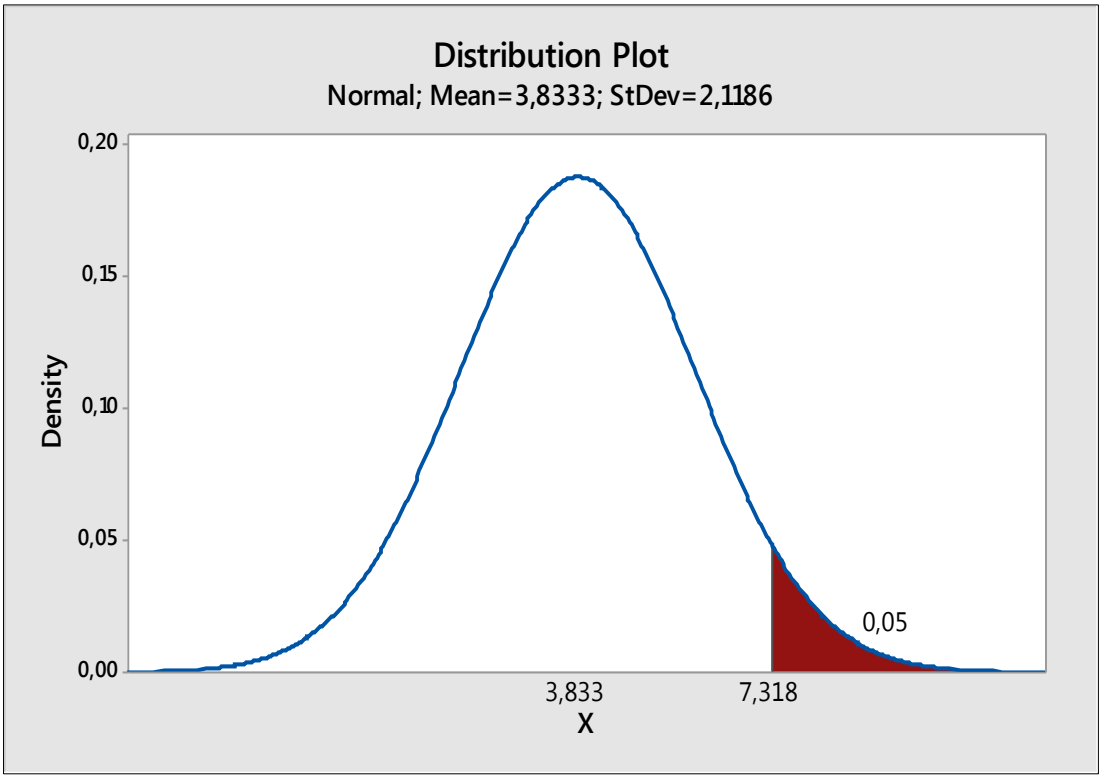


*Figura 80*. Distribución 5

## c) Cálculo: Prueba t para prueba de medias de las dos muestras

Tabla 47
*Tiempo para procesar información del KPI5*

| | **Variable 1** | **Variable 2** |
|---|---|---|
| Media | 23,2 | 3,8333333 |
| Varianza | 11,9586207 | 4,48850575 |
| Observaciones | 30 | 30 |
| Diferencia hipotética de las medias | 0 | |
| Grados de libertad | 29 | |
| Estadístico t | 24,42,4972 | |
| P(T<=t) dos colas | 6,8439E-21 | |
| Valor critico de t(dos colas) | 2,04522964 | |

**d) Decisión Estadística**

Puesto que el valor-p=0.000 < α=0.05, los resultados proporcionan suficiente evidencia para rechazar la hipótesis nula (Ho), y la hipótesis alterna (Ha) es cierta. La prueba resultó ser significativa.

### 4.3.6 Prueba para el indicador: Tiempo para generar exactitud KPI6

Se valida el impacto que tiene al realizar un modelo predictivo en el Tiempo para generar exactitud, llevado a cabo en la muestra. Se realizó una medición antes de realizar el modelo predictivo (Post-Prueba O2) y otra después de realizar un modelo predictivo (Post-Prueba O1).

Hipótesis Específica Hi6: Si se realizar el modelo predictivo para el proceso de refinanciamiento de los alumnos en la Universidad Autónoma del Perú, se disminuirá el Tiempo para generar exactitud (Post-Prueba O1) con respecto a la muestra a la cual se realiza (Post-Prueba O2).

Tabla 48

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 8 | 9 | 8 | 7 | 6 | 10 | 6 |
| | 7 | 6 | 5 | 5 | 8 | 8 | 10 |
| **Pre - Prueba** | 6 | 7 | 6 | 5 | 7 | 6 | 9 |
| | 8 | 7 | 6 | 9 | 8 | 7 | 15 |
| | 18 | 20 | | | | | |

*Prueba Kpi6*

Tabla 49

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 5 | 4 | 5 | 7 | 8 | 6 | 5 |
| | 4 | 3 | 3 | 4 | 2 | 4 | 4 |
| **Pre - Prueba** | 2 | 1 | 8 | 7 | 6 | 9 | 8 |
| | 8 | 8 | 8 | 7 | 6 | 9 | 8 |
| | 8 | 8 | | | | | |

*Prueba Kpi6*

**Hi6:** El uso de un modelo predictivo disminuye el Tiempo para generar exactitud **(Post-Prueba)** con respecto a la muestra a la que no se aplicó **(Pre-Prueba).**

**Solución:**

**a) Planteamiento de la Hipótesis**

**µ1=** Media del Tiempo para registrar las incidencias Pre-Prueba.

**µ2=** Media del Tiempo para registrar las incidencias Post-Prueba

$$H_o: \mu_1 \leq \mu_2$$
$$H_a: \mu_1 > \mu_2$$

**b) Criterio de Decisión:**

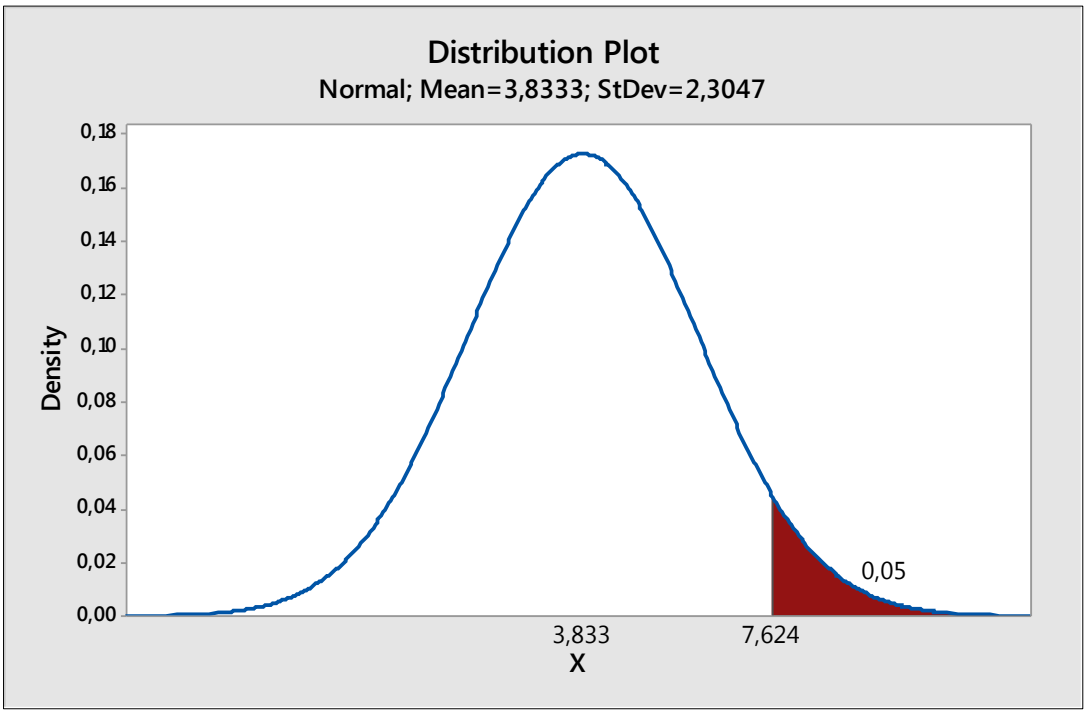


*Figura 81*. Distribución 6

## c) Cálculo: Prueba t para prueba de medias de las dos muestras

Tabla 50
*Tiempo para generar exactitud del KPI6.*

| | *Variable 1* | *Variable 2* |
|---|---|---|
| Media | 8,23333333 | 5,83333333 |
| Varianza | 12,5298851 | 5,31609195 |
| Observaciones | 30 | 30 |
| Coeficiente de correlación de Pearson | 0,28378276 | |
| Diferencia hipotética de las medias | 0 | |
| Grados de libertad | 29 | |
| Estadístico t | 3,61624726 | |
| P(T<=t) una cola | 0,00056075 | |
| Valor crítico de t (una cola) | 1,69912703 | |
| P(T<=t) dos colas | 0,00112149 | |
| Valor crítico de t (dos colas) | 2,04522964 | |

## e) Decisión Estadística

Puesto que el valor-p=0.000 < α=0.05, los resultados proporcionan suficiente evidencia para rechazar la hipótesis nula (Ho), y la hipótesis alterna (Ha) es cierta. La prueba resultó ser significativa.

# CAPÍTULO V
# CONCLUSIONES Y RECOMENDACIONES

## 5.1 CONCLUSIONES

- Después de la revisión de la literatura podemos afirmar que el tema de investigación realizado es un tema abierto a la polémica y discusión.
- Se comprueba que el modelo predictivo, tiene un alto porcentaje de confiabilidad, al dar los resultados con altos porcentajes de aseveración.
- Se comprueba que el modelo clasifica los datos para poder predecir a los posibles alumnos morosos.
- Se aprecia que el estudio realizado ha dado los valores cualitativos necesarios para hacer la minería de datos.
- Se observa que el modelo predictivo se pudo generar mediante los diferentes datos recogidos y puestos en funcionamiento como conocimiento y entrenamiento para mejorar el nivel de predicción del modelo ya descrito.
- Los beneficios obtenidos gracias al modelo predictivo, entre las cosas que se han obtenido ha sido la reducción del tiempo y la acertividad de si un alumno será moroso o no.
- Se comprueba que los métodos de minería de datos para ejecutar modelos predictivos, tienen un alto índice de fiabilidad según sea el modelo usado y para que esta destinado.

## 5.2 RECOMENDACIONES

- Se sugiere, continuar implementado la metodología CRISP-DM, para el desarrollo de un modelo predictivo con mayor cobertura. Ya que es una metodología de fácil entendimiento y nos detalla el paso a paso cómo seguir el desarrollo de un modelo predictivo.
- Se sugiere, obtener mas apoyo del centro donde se realizo el estudio para poder dar con mejor calidad la información de la predicción si los estudiantes serán morosos o no.
- Se aconseja, investigar en el mercado sobre las nuevas tendencias tecnológicas que tienen relación con la toma de decisiones para los procesos de minería de datos, ya que de esta manera se iría mejorando el modelo hasta hacer uno más autónomo.
- Se aconseja realizar un plan de contingencia pos-implementación, para darle mayor continuidad y calidad al modelo predictivo.

# REFERENCIAS BIBLIOGRÁFICAS

**Tesis**

Fischer, E. (2012). *Modelo para la automatización del Proceso de Determinación de Riesgo de Deserción en estudiantes universitarios* (Tesis de maestría). Recuperado de http://repositorio.uchile.cl/bitstream/handle/2250/111188/cf-fischer_ea.pdf?sequence=1&isAllowed=y

Moreno, M y Ovalle, V. (2011). *Aplicación de un modelo predictivo de fuga de clientes utilizando Data Mining en VTR Globalcom S.A Zona Sur* (Tesis de pregrado). Recuperado de http://repobib.ubiobio.cl/jspui/bitstream/123456789/2308/1/Ovalle_Retamal_Victor_Francisco_Javier.pdf

Salinas, J. (2005). *Reconocimiento de Patrones de Morosidad para un Producto Crediticio usando la Técnica de Árbol de Clasificación Cart*. (Tesis de maestría). Recuperado de http://cybertesis.unmsm.edu.pe/bitstream/handle/cybertesis/3108/Salinas_fj.pdf?sequence=1&isAllowed=y

**Artículo**

Chuquival, J., Galindo Jaime y Maquera, S. (2012). Modelo de redes neuronales para mejorar el pronóstico del comportamiento del alumno en el cumplimiento del pago de sus armadas, concernientes a un crédito aprobado por el área de finanzas Alumnos de la Universidad Peruana Unión. *Busisness Intelligence, 1*(2), 12-17. Recuperado de https://revistas.upeu.edu.pe/index.php/ri_bi/article/view/914

Mallo, P. (2010). Análisis de la Morosidad Tributaria de las Empresas Aplicando Técnicas Borrosas y Estadísticas. El Caso de Mar del Plata. *Anales de las Jornadas Internacionales de Estadístic*a, *1*(1), 1-9. Recuperado de https://bit.ly/2HgauZq

Vargas, H., Ccapa, L. (2011). Modelo de Árboles de decisión para pronosticar la morosidad de los alumnos de la Universidad Peruana Unión. *Busisness Intelligence*, *1*(2), 26-32. Recuperado de https://revistas.upeu.edu.pe/index.php/ri_bi/article/view/916

**Sitios Web**

Goicochea, A. (11 de agosto de 2012). Aníbal Goicochea. Recuperado de https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectosde-mineria-de-datos/

Google Maps. (10 de Julio de 2016). Ubicación de la Universidad Autónoma del Perú. Recuperado de https://www.google.com.pe/maps/place/Universidad+Aut%C3%B3noma+del+Per%C3 %BA/@-12.1955257, 76.9739164, 17z/data=!3m1!4b1!4m5!3m4!1s0x9105b9989f7875ef:0xac24a8fceeb849 a!8m2!3d-12.1955257!4d-76.9717277

Gutiérrez, J. (14 de marzo de 2013). *Metereologia de Santander*. Recuperado de http://www.meteo.unican.es/es/research/mineria_datos

Inacap. (15 de abril de 2014). Inacap. Recuperado de *ht*tp://inacap.serveftp.com/tic2/Presentaciones-N2/Algoritmo%20de%20Clasificaci%C3%B3n%202%C2%B0%20Informe.pdf

Inei. (12 de octubre de 2012). *Inei*. Recuperado de http://censos.inei.gob.pe/cenaun/redatam_inei/doc/ESTADISTICA_UNIVERSITARIAS.pdf

Nivel, A. (14 de agosto de 2014). *ALto Nivel*. Recuperado de http://www.altonivel.com.mx/36690-arbol-de-decision-una-herramienta-para-decidircorrectamente.html

Pablo, J. (2 de febrero de 2012). *Metodologia SEMMA*. Recuperado de http://elbuhoanaltico.blogspot.pe/2012/02/metodologia-semma.html

Pete, C. (13 de Setiembre de 2011). *Data Prix*. Recuperado de http://www.dataprix.com/es/metodologcrisp-dm-para-miner-datos

Rodríguez, O. (24 de febrero de 2014). Crisp – DM. Recuperado de http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISPDM.2385037.pdf

Universidad Autónoma del Perú. (10 de marzo de 2016). Organigrama de la Universidad Autónoma del Perú. Recuperado de http:// www.autonoma.com

UTM. (11 de agosto de 2012). Árbol de decisión. Recuperado de *http://www.utm.mx/~jahdezp/archivos%20estructuras/DESICION.pdf*

# ANEXOS Y APÉNDICES

## Apéndice 1. Matriz de consistencia

| | | |
|---|---|---|
| **TITULO** | APLICACIÓN DE MINERIA DE DATOS PARA PRONOSTICAR EL RIESGO DE MOROSIDAD DE LOS ESTUDIANTES DE LA UNIVERSIDAD AUTONOMA DEL PERU | |
| **PROBLEMA** | ¿DE QUE MANERA LA APLICACIÓN DE UN MODELO PREDICTIVO BASADO EN LA METODOLOGIA CRISP-DM NOS AYUDARA A PREDECIR DE MEJOR MANERA A LOS ALUMNOS MOROSOS DE LA UNIVERSIDAD AUTONOMA DEL PERU? | |
| **OBJETIVOS** | GENERAL | Determinar en qué medida el uso de una aplicación de minería de datos influenciara en el pronóstico de riesgo de morosidad de los estudiantes de la Universidad Autónoma del Peru-2016. |
| | ESPECIFICO | Determinar en qué medida un aplicativo de minería de datos basado en la técnica de árboles reducirá el tiempo para verificar el alumno moroso de la Universidad Autónoma del Perú en el año 2016. |
| | | Determinar en qué tiempo un modelo predictivo clasificaría a los posibles alumnos morosos de la Universidad Autónoma del Perú en el año 2016 |
| | | Determinar en qué medida un aplicativo de minería de datos basado en la técnica de árboles reducirá el tiempo para verificar el alumno moroso de la Universidad Autónoma del Perú en el año 2016. |
| | | Determinar en qué tiempo se realizaría el estudio cualitativo del comportamiento de las personas de la Universidad Autónoma del Perú en el año 2016. |
| | | Determinar una investigación del comportamiento de las personas para generar conocimiento sobre los riesgos en los otorgamientos de facilidades de la Universidad Autónoma del Perú en el año 2016. |
| | | Determinar los beneficios que se obtendrán al aplicar el modelo predictivo en la Universidad Autónoma del Perú en el año 2016. |

| | | |
|---|---|---|
| | | Determinar los beneficios de usar Data Mining para la obtención del modelo predictivo en la Universidad Autónoma del Perú en el año 2016. |
| **HIPOTESIS** | El uso de un modelo predictivo basado en la técnica de árboles influye en la predicción de riesgo al otorgar facilidades económicas a los estudiantes de la Universidad Autónoma del Perú | |
| **VARIABLES** | INDEPENDIENTE | Modelo predictivo usando Data Mining para poder clasificar. |
| | INTERVINIENTE | Metodología CRISP-DM |
| | DEPENDIENTE | Predicción de morosidad |
| **INDICADORES** | VARIABLE INDEPENDIENTE | Presencia - Ausencia |
| | VARIABLE DEPENDIENTE | Tiempo para verificar datos del alumno |
| | | Tiempo para verificar alumno moroso |
| | | Índice de Riesgo |
| | | Tiempo para Determinar el riesgo de morosidad de un estudiante |
| | | Tiempo para predecir que alumnos incurrirán en morosidad |
| **TIPO Y DISEÑO DE LA INVESTIGACION** | Tipo de Investigación: Aplicada. Nivel de Investigación: - Explicativa- Experimental - Correlacional. Diseño de la Investigación: Gc O1 X O2 | |

Anexo 1

Encuesta

**ENTREVISTA**

1. ¿Quién solventa tus estudios?

a) Tu b) Padres c) Otros

2. ¿Selecciona tu sexo?

a) Hombre b) Mujer

3. ¿Con que frecuencia ahorras?

a) Mensual b) Semanal c) Diario

4. ¿De qué manera sustentas un préstamo?

a) Aval b) Empeño c) Hipoteca

5. ¿Alguna vez requirió ayuda en la asistenta social?

a) Si b) No

6. ¿Qué tipo de vivienda tiene?

a) Propia b) Alquilada c) Otros

7. ¿Cuántos hermanos tienes actualmente?

a) 1 b) 2 c) 3 d)más de 3

8. ¿Cuál es el ingreso mensual en casa?

a) 1000 – 2000 b) 2000 – 3000 c) Mas de 3000

9. ¿Acostumbra salir todos los fines de semana?

a) Si b) No

10. ¿Cuánto es su pago mensual en la Universidad?

a) 280 – 350 b) 350 – 450 c) 450 – 550 d) más de 600

# GLOSARIO DE TÉRMINOS

A

- Asimetría: Esta medida nos permite identificar si los datos se distribuyen de forma uniforme alrededor del punto central (Media aritmética). La asimetría presenta tres estados diferentes, cada uno de los cuales define de forma concisa como están distribuidos los datos respecto al eje de asimetría. Se dice que la asimetría es positiva cuando la mayoría de los datos se encuentran por encima del valor de la media aritmética, la curva es Simétrica cuando se distribuyen aproximadamente la misma cantidad de valores en ambos lados de la media y se conoce como asimetría negativa cuando la mayor cantidad de datos se aglomeran en los valores menores que la media.

B

- Biblioteca: Es un lugar en donde se almacenan libros que por su organización facilita la búsqueda de una información determinada.

C

- Cadena de valor: Es un modelo teórico que permite describir el desarrollo de las actividades de una organización empresarial generando valor al cliente final.

- Cartera de negocios: Es el conjunto de los negocios y productos que forman una compañía.

- Cronograma de producción: Recoge las actividades realizadas en el equipo de desarrollo durante la iteración.

- CSS: Las hojas de estilo en cascada o hacen referencia a un lenguaje de hojas de estilos usado para describir la presentación semántica de un documento escrito en lenguaje de marcas.

H

- Hipótesis alternativa: Conocida como, es cualquier hipótesis que difiere de la hipótesis nula.

- Hipótesis nula: Se denomina hipótesis nula a la hipótesis que se desea contrastar.

- Historia de usuario: Una historia de usuario (o user history en Inglés) describe una funcionalidad que, por sí misma, aporta valor al usuario.

- HTML: Son las siglas de HyperText Markup Language, hace referencia al lenguaje de marcado para la elaboración de páginas web. Este estándar de la define una estructura básica y un código para el desarrollo de contenido de una página web, texto como, imágenes, etc.

I

- Indicador: Magnitud utilizada para medir o comparar los resultados efectivamente obtenidos, en la ejecución de un proyecto, programa o actividad. Resultado cuantitativo de comparar dos variables.

K

- KPI: En inglés Key Performance Indicators, o Indicadores Clave de Desempeño, Miden el nivel del desempeño de proceso, centrándose en el "como" e indicando el rendimiento de los procesos, de forma que se pueda alcanzar el objetivo fijado.

- Kurtosis: Esta medida determina el grado de concentración que presentan los valores en la región central de la distribución. Por medio del Coeficiente de Curtosis, podemos identificar si existe una gran concentración de valores (Leptocúrtica), una concentración normal (Mesocúrtica) o una baja concentración (Platicúrtica).

M

- Minitab: Es un software diseñado para ejecutar funciones estadísticas básicas y avanzadas.

- Muestra (estadística): Subconjunto de los individuos de una población estadística. Una muestra permite inferir las propiedades del total del conjunto.

N

- Nivel de significación (σ): Se define como la probabilidad de rechazar erróneamente la hipótesis nula.

S

- Smartphone (teléfono inteligente): Móvil que ofrece servicios propios de un ordenador. Para ello, suele tener un sistema operativo avanzado (iOS, Android, Windows, etc) que le permite acceder a Internet, servicios de email, organizador personal, descarga de aplicaciones, etc.

- Sprint: Es el período en el cual se lleva a cabo el trabajo en sí.

- Stakeholder: Son personas u organizaciones (por ejemplo, clientes, patrocinadores, la organización ejecutante o el público), que participan activamente en el proyecto, o cuyos intereses pueden verse afectados positiva o negativamente por la ejecución o terminación del proyecto.

- SXP: La metodología ágil SXP es la unión de XP y SCRUM. Desarrollada en 2007 en la Universidad de las Ciencias Informáticas está indicada especialmente para proyectos pequeños.

T

- T Test (2 sample): Prueba que usa un test de hipótesis para las medias de dos poblaciones con el fin de determinar si son significativamente diferentes.

V

- Valor de P: Conocido como la P, P-Valor o P-Value por su traducción al inglés, determina la conveniencia de rechazar la hipótesis nula de una prueba de la hipótesis.

- Variable dependiente: Es el factor que el investigador observa o mide para determinar el efecto de la variable o variable causa. La variable dependiente es la variable respuesta o variable salida u output. A la variable dependiente se le considera así porque sus valores van a depender de los valores de la variable independiente.

- Variable independiente: Es la variable que el investigador mide, manipula o selecciona para determinar su relación con el fenómeno o fenómenos observados. Esta variable es conocida también como variable estímulo o input.

- Variable interviniente: Son aquellas que teóricamente afectan a la variable dependiente pero no pueden medirse o manipularse. Normalmente son variables que se deducen de los efectos de las variables: independiente y moderador, sobre la variable dependiente.