FACULTAD DE INGENIERÍA

CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

"APLICACIÓN DE MINERÍA DE DATOS BASADO EN ÁRBOLES DE DECISIÓN PARA PREDECIR EL RIESGO DE MOROSIDAD DE LOS CLIENTES EN LA EMPRESA DE SEGUROS ONCOSALUD S.A.C. 2018"

PARA OBTENER EL TÍTULO DE

INGENIERO DE SISTEMAS

AUTOR(ES)

LEONARDO ESPINO QUIÑONES

MARÍA EMILY GARCÍA TORRES

ASESOR

ING. RAMÓN JOHNY PRETELL CRUZADO

LIMA, PERÚ, DICIEMBRE DE 2018

DEDICATORIA

En especial a mi padre celestial por permitirme llegar en este momento tan especial en mi vida, a mis padres por su cariño y confianza, y por estar siempre conmigo brindándome su apoyo incondicional en todo momento para lograr mis objetivos como profesionales.

García Torres María Emily

A mis padres por ser quienes me enseñaron el valor de luchar día a día para conseguir mis metas, a mi padre celestial por permitirme llegar a este momento tan especial porque es el que me da la vida y fortaleza para desarrollarme completamente en todos los aspectos de mi vida.

Espino Quiñones Leonardo

AGRADECIMIENTOS

A la Universidad Autónoma del Perú por formarnos y desarrollarnos como profesionales brindándonos aportes invaluables que nos servirán en nuestras vidas. Así como también a todos los docentes que brindaron su apoyo para seguir día a día para el desarrollo de nuestra tesis apoyándonos en todo momento para poder lograr que este sueño se haga realidad.

Agradecemos también al director y al personal del departamento de cobranza de la empresa de seguros Oncosalud por habernos aceptado que se realice el proyecto de investigación y por todo el apoyo y facilidades que nos fueron otorgadas en la empresa. Por darnos la oportunidad de crecer profesionalmente.

Damos gracias a nuestro asesor quien nos orientó en todo momento en nuestro proyecto de investigación, transmitiendo sus conocimientos, su manera de trabajar, su persistencia, su paciencia que han sido fundamentales para nuestra formación como investigadores y a su vez inculcando en nosotros un sentido de seriedad, responsabilidad y rigor académico.

Los Autores.

RESUMEN

En el Perú y en distintos lugares del mundo la morosidad se ha convertido en un serio problema para las pequeñas, medianas y grandes empresas. Estas empresas cuentan con un gran manejo de clientes y muchos de ellos incumplen con el plazo pactado en sus pagos, esto puede traer consecuencias de diversos tipos hasta verse obligado a cerrar la empresa por la imposibilidad económica por parte de los clientes. Es importante mencionar que la morosidad no solo afecta al sector bancario sino también a distintos sectores como por ejemplo en el mercado de seguros de salud.

La empresa Oncosalud se encarga de brindar seguros Oncológicos contando con un gran volumen de clientes que acuden a este servicio para su tratamiento correspondiente. Sin embargo, el departamento de cobranza identificó que muchos de los clientes asegurados están incumpliendo con sus pagos en la fecha pactada, de manera que se está generando alto índices de morosidad.

Para esta problemática se desarrolló un modelo predictivo, utilizando la técnica minería de datos basado en árboles de decisión utilizando el algoritmo ID3 con el objetivo de facilitar la predicción del riesgo de morosidad al momento que se le haya vendido un seguro oncológico a un cliente, con el fin de aplicar estrategias anticipadas en el departamento de cobranza y tomar decisiones adecuadas con mayor grado certeza utilizando información histórica de los clientes.

En este estudio se aplicó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) comprendida en sus 6 fases, por lo que se desarrolló paso a paso utilizando el software Weka tomando las variables que presentaron mayor relevancia en el estudio para el desarrollo del modelo. Finalmente, para mostrar los resultados y a su vez facilitar la predicción de morosidad de los clientes se desarrolló un sistema Web integrado con Weka desarrollado en JSP (Java Server Page).

Palabras clave: Minería de datos, predicción del riesgo de morosidad, metodología CRISP-DM, arboles de decisión.

ABSTRACT

In Peru and in different parts of the world, delinquency has become a serious problem for small, medium and large companies. These companies have a large customer management and many of them fail to meet the deadline agreed on their payments, this can have consequences of various types to be forced to close the company because of the economic impossibility by customers. It is important to mention that delinquency not only affects the banking sector but also in different sectors such as, for example, in the health insurance market.

The Oncosalud Company is responsible for providing Oncological insurance with a large volume of clients who come to this service for their corresponding treatment. However, the collection department identified that many of the insured clients are not complying with their payments on the agreed date, so that high delinquency rates are being generated.

For this problem, a predictive model was developed, using the data mining technique based on decision trees using the ID3 algorithm with the aim of facilitating the prediction of the risk of delinquency at the time that an oncological insurance was sold to a client. In order to apply anticipated strategies in the collection department and make appropriate decisions with greater certainty degree using historical information of the clients.

In this study, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was applied, comprised in its 6 phases, so it was developed step by step using the Weka software taking the variables that were most relevant in the study for the development of the model. Finally, to show the results and in turn to facilitate the prediction of customer delinquency, an integrated Web system was developed with Weka developed in JSP (Java Server Page).

Keywords: Data mining, prediction of delinquency risk, methodology CRISP-DM, decision trees.

ÍNDICE DE CONTENIDO

DEDICATORIA i
AGRADECIMIENTOS ii
RESUMEN iii
ABSTRACT iv
INTRODUCCIÓN xii

CAPÍTULO I. PLANTEAMIENTO METODOLÓGICO

1.1. EL PROBLEMA 2
1.1.1. Descripción de la Realidad Problemática 2
1.1.2. Definición del problema 5
1.1.3. Enunciado del problema 10
1.2. TIPO Y NIVEL DE LA INVESTIGACIÓN 10
1.2.1. Tipo de Investigación 10
1.2.2. Nivel de Investigación 10
1.3. JUSTIFICACIÓN DE LA INVESTIGACIÓN 10
1.4. OBJETIVOS DE LA INVESTIGACIÓN 11
1.4.1. Objetivo general 11
1.4.2. Objetivos específicos 11
1.5. HIPÓTESIS 12
1.6. VARIABLES E INDICADORES 12
1.6.1. Variable Independiente 12
1.6.2. Variable Dependiente 12
1.7. LIMITACIONES DE LA INVESTIGACIÓN 13
1.8. DISEÑO DE LA INVESTIGACIÓN 14
1.9. TÉCNICAS E INSTRUMENTO PARA LA RECOLECCIÓN DE INFORMACIÓN. 15
1.9.1. Técnicas e Instrumentos 15

CAPÍTULO II. MARCO TEÓRICO

2.1. ANTECEDENTES DE ESTUDIOS 18

2.2. DESARROLLO DE LA TEMÁTICA CORRESPONDIENTE AL TEMA INVESTIGADO 25

2.2.1. Riesgo de morosidad 25

2.2.1.1. Morosidad 25

2.2.1.2. Características de la morosidad 25

2.2.1.3. Causas de la morosidad 26

2.2.1.4. Tipos de morosos 27

2.2.1.5. Clasificación por categoría de riesgo 28

2.2.1.6. Dimensiones de la predicción del riesgo de morosidad 29

2.2.2. Minería de Datos 30

2.2.2.1. Tareas y técnicas de la minería de datos. 31

2.2.2.2. Análisis predictivo 33

2.2.2.3. Árboles de decisión 34

2.2.2.4. Algoritmos para árboles de decisión: 37

2.2.2.5. Metodologías para implementar Minería de datos 40

2.3. DEFINICIÓN CONCEPTUAL DE LA TERMINOLOGÍA EMPLEADA 47

2.4. ESTADO DEL ARTE 48

CAPÍTULO III. DESARROLLO DEL SISTEMA

3.1. ESTUDIO DE FACTIBILIDAD 54

3.1.1. Factibilidad técnica 54

3.1.2. Factibilidad operativa 54

3.1.3. Factibilidad económica 55

3.2. APLICACIÓN DE LA METODOLOGÍA CRISP-DM 56

3.2.1. Comprensión del Negocio 56

3.2.2. Comprensión de los datos 60

3.2.3. Preparación de los datos 66

3.2.4. Modelado 71

3.2.5. Evaluación 85

3.2.6. Implantación 86

CAPITULO IV. ANÁLISIS DE RESULTADOS Y CONTRASTACIÓN DE LA HIPÓTESIS

4.1 POBLACIÓN Y MUESTRA 95

4.1.1 Población 95

4.1.2 Muestra 95

4.2 VALIDEZ Y CONFIABILIDAD DEL INSTRUMENTO 96

4.2.1 Validez 96

4.2.2 Confiabilidad 96

4.2 ANÁLISIS E INTERPRETACIÓN DE RESULTADOS 97

4.2.1 Resultados 97

4.3 NIVEL DE CONFIANZA Y GRADO DE SIGNIFICANCIA 105

4.4 PRUEBA DE HIPÓTESIS 105

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES 113

5.2 RECOMENDACIONES 114

REFERENCIAS BIBLIOGRÁFICAS

ANEXOS Y APÉNDICES

ÍNDICE DE TABLAS

Tabla 1 Conceptualización de la variable Independiente 12
Tabla 2 Indicador Variable Independiente 12
Tabla 3 Conceptualizacion de indicadores de la variable dependiente 13
Tabla 4 Indicador Variable Dependiente 13
Tabla 5 Diseño de la investigación 14
Tabla 6 Técnicas e instrumentos de la investigación 16
Tabla 7 Cuadro comparativo de algoritmos de minería de datos 40
Tabla 8 Comparación entre las metodologías KDD, CRISP-DM, SEMMA y CATALYST 46
Tabla 9 Software disponible 54
Tabla 10 Costos de desarrollo de la solución 55
Tabla 11 Indicadores para medir el criterio de éxito del negocio 58
Tabla 12 Plan del proyecto 60
Tabla 13 Input y output de la recolección de los datos 61
Tabla 14 Diccionario de datos 62
Tabla 15 Variables objetivo del modelo 68
Tabla 16 Selección de la técnica de modelado 72
Tabla 17 Matriz de confusión 72
Tabla 18 Ganancia de información de variables 76
Tabla 19 Cálculo de la entropía 77
Tabla 20 Ganancia de información de variables 78
Tabla 21 Entropia de variables 79
Tabla 22 Tercera parte de la ganancia de información 79
Tabla 23 Entropia de variables 80
Tabla 24 Ganancia de información 81
Tabla 25 Resultado de la matriz de confusión del modelo desarrollado 85
Tabla 26 Validez de los instrumentos por expertos 96
Tabla 27 Matriz de confusión para la Pre prueba del KPI1 y KPI2 97
Tabla 28 Resultado de la Pre-Prueba y Post-Prueba para el KPI1 98
Tabla 29 Resultado de la Pre-prueba y Post-Prueba para el KPI2 99
Tabla 30 Resultado de la Pre-Prueba para el KPI3 100

Tabla 31 Frecuencia de la Pre-Prueba para el KPI3 100
Tabla 32 Resultado de la Post-Prueba para el KPI3 101
Tabla 33 Frecuencia de la Post-prueba para el KPI3……....102
Tabla 34 Comparación de los tiempos promedios del KPI4 104
Tabla 35 Indicadores para la Contrastación de la hipótesis.....105
Tabla 36 Estadística de muestras relacionadas del indicador nivel de dificultad KPI3..... 108
Tabla 37 Resultados de la prueba t student para el KPI3 108
Tabla 38 Estadísticos descriptivos para el KPI4..... 110

ÍNDICE DE FIGURAS

Figura 1 Ratio de morosidad del mundo. 3
Figura 2 Ratio de morosidad en el Perú 4
Figura 3 Ubicación de la empresa Oncosalud.. 5
Figura 4 Monto no cobrado de clientes de la empresa Oncosalud 6
Figura 5 Índice de morosidad entre los meses enero y abril 2018 6
Figura 6 Proceso de cobranza de la empresa Oncosalud 7
Figura 7 Proceso de ventas de seguro Oncológicos. 8
Figura 8 Proceso de evaluación con el sistema implementado. 9
Figura 9 Tareas de la minería de datos. 31
Figura 10 Estructura de un árbol de decisión. 34
Figura 11 Grafo. 39
Figura 12 Fases de la metodología KDD. 41
Figura 13 Fases de la metodología CRISP-DM.. 42
Figura 14 Fases de metodología SEMMA. 44
Figura 15 Interacción de los diferentes modelos (MII – MIII) 45
Figura 16 Metodologías más usadas. 47
Figura 17 Ejemplo de predicción de riesgo de morosidad basado en la terminología. 50
Figura 18 Situación actual del negocio 57
Figura 19 Porcentaje de clientes por riesgo de morosidad. 64
Figura 20 Cantidad de clientes por categoría y riesgo de morosidad. 64
Figura 21 Cantidad de cliente según su sexo por riesgo de morosidad. 65
Figura 22 Cantidad de clientes por programa. 65
Figura 23 Cantidad de clientes con su calificación crediticia mediante el riesgo. 66
Figura 24 Datos preparados para el modelo. 71
Figura 25 Data de los clientes de la empresa. 74
Figura 26 Primera parte del árbol de decisión. 77
Figura 27 Segunda parte del árbol de decisión. 78
Figura 28 Tercera parte del árbol de decisión. 80
Figura 29 Modelo de árbol de decisión completo. 81

Figura 30 Datos en Weka. 82
Figura 31 Selección del algoritmo ID3 en Weka. 82
Figura 32 Resultados del algoritmo ID3. 83
Figura 33 Matriz de confusión del modelo. 84
Figura 34 Modelo de la arquitectura del sistema 87
Figura 35 Pantalla de inicio de sesión del sistema. 89
Figura 36 Listado de clientes en el sistema. 89
Figura 37 Resultado del algoritmo ID3 generado por el sistema 90
Figura 38 Registro de nuevos clientes del sistema. 90
Figura 39 Filtración de datos del sistema. 91
Figura 40 Gráficos de reportes del sistema. 91
Figura 41 Registro de un cliente en el sistema. 92
Figura 42 Resultado de la predicción de un cliente en el sistema 92
Figura 43 Validez del indicador Nivel de dificultad. 96
Figura 44 Validez del indicador Tiempo para predecir. 97
Figura 45 Precisión de la predicción de la Pre-Prueba y Post-Prueba. 98
Figura 46 Error de predicción de la Pre-Prueba y Post-Prueba. 99
Figura 47 Nivel de dificultad Pre-prueba. 101
Figura 48 Nivel de dificultad para la Post-Prueba. 102
Figura 49 Resultado del tiempo para predecir (KPI4) en la Pre-Prueba. 102
Figura 50 Resultado del tiempo para predecir (KPI4) en la Post-Prueba. 103
Figura 51 Gráfico de distribución KPI1 107
Figura 52 Gráfico de distribución para I3. 109
Figura 53 Gráfico de distribución para el I4. 111

INTRODUCCIÓN

La morosidad es uno de los factores más importantes que deben combatirse a nivel mundial, hoy en día reducir el impacto de la morosidad es uno de los principales objetivos de las pequeñas, medianas y grandes empresas, causadas por los pagos impuntuales de los clientes.

Es importante mencionar que hoy por hoy es habitual que las empresas de distintos sectores tomen decisiones a diario, tomando en consideración que el éxito o fracaso pueden depender de cada determinación, por esta razón debemos apoyarnos de herramientas y tecnologías que me ayuden a tomar decisiones apropiadas para conseguir el resultado deseado.

Continuando bajo la misma perspectiva, el proyecto de investigación consiste en la implementación de un modelo predictivo utilizando técnicas de minería de datos desarrollado bajo la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), para mejorar principalmente en la toma de decisiones en el departamento de cobranza de la empresa de seguro Oncosalud.

El proyecto de investigación está aplicado bajo la técnica de clasificación, representadas esquemáticamente en árboles de decisión de tal manera que la decisión final a tomar se determina mediante lo que muestra el árbol de decisión.

Los resultados de esta investigación serán mostrados en un sistema, para ser evaluado bajo los analistas del departamento de cobranza con el fin de mostrar los resultados de manera más interpretativa.

Para comprender un poco más respecto al proyecto de estudio, se realiza la división de esta investigación por capítulos, donde se explica a continuación:

Capítulo I: Planteamiento Metodológico. - En este capítulo se especifica todo referente al planeamiento metodológico donde se define el problema, tipo y nivel de investigación, justificación de la investigación, objetivos de la investigación, hipótesis, variables e indicadores, limitaciones de la investigación, diseño de la investigación y las técnicas e instrumentos para la recolección de información.

Capítulo II: Marco Teórico. - En este capítulo se detallan los antecedentes de estudio que se relación con la investigación, de igual manera, artículos

científicos, se conceptualiza las variables dependientes, independiente e intervinientes para estructurar el estudio que se desarrolla.

Capítulo III: Desarrollo del sistema. - En este capítulo en primer lugar se ha descrito la factibilidad de la investigación y se ha desarrollado la metodología que se ha seleccionado para esta investigación el cual ha sido la metodología de minería de datos CRISP-DM siguiente sus seis fases para su desarrollo.

Capítulo IV: Análisis de resultados y contrastación de la hipótesis. - En este capítulo se describió la población y muestra de la investigación. También se analizó los instrumentos de la investigación, tanto la validez por expertos y la confiabilidad. Se analizó los resultados de los indicadores mediante pruebas estadísticas, seguidamente se realizó la contratación de hipótesis.

Capítulo V: Conclusiones y recomendaciones. - En este último capítulo se detalla las conclusiones y recomendaciones de la investigación.

Al final se presentan las referencias bibliográficas, anexos y apéndices.

CAPÍTULO I
PLANTEAMIENTO METODOLÓGICO

1.1. EL PROBLEMA

1.1.1. Descripción de la Realidad Problemática

Realidad Mundial

La situación social que atraviesa la economía internacional y la imposibilidad de efectuar pagos en la fecha fijada es muy recurrente, lo que origina la morosidad de cientos de clientes afiliadas a una empresa, la morosidad se trata de un fenómeno que ha estado y seguirá estando presente en las cuentas de la mayoría de las empresas de diversos sectores en el mundo.

Según el informe facilitado por el DAS (Compañía de protección jurídica) La unión europea se encuentra frente a alarmantes cifras de impagos especialmente en los países del sur, entre el año 2008 y 2014 la morosidad provocó el cierre nada menos que de 400,000 empresas en Europa y en el año 2015 las pymes en España tenían facturas de un 44% fuera de plazo pendientes por cobrar, esto evidencia que la morosidad siguen siendo un motivo de preocupación importante para las pymes en la Unión Europea.

Los países de la Unión Europea disponían de dos años para modificar sus leyes y adaptarse así al nuevo marco de la legislación europea para poder acabar con la cultura del impago, pero a pesar de estos intentos normativos los problemas de morosidad persisten y no ha surgido el mismo efecto que en otros lugares, lo que sitúa a España superado por Portugal, Italia y Reino Unido con mayor retraso en los pagos.

Esta morosidad provoca que en el día a día las empresas vayan perdiendo capacidad para continuar con sus inversiones y pagos, hasta llegar al punto que su empresa caiga por completo en su actividad financiera y genere el cierre de la organización.

Europa es la región con mayor ratio de morosidad del mundo

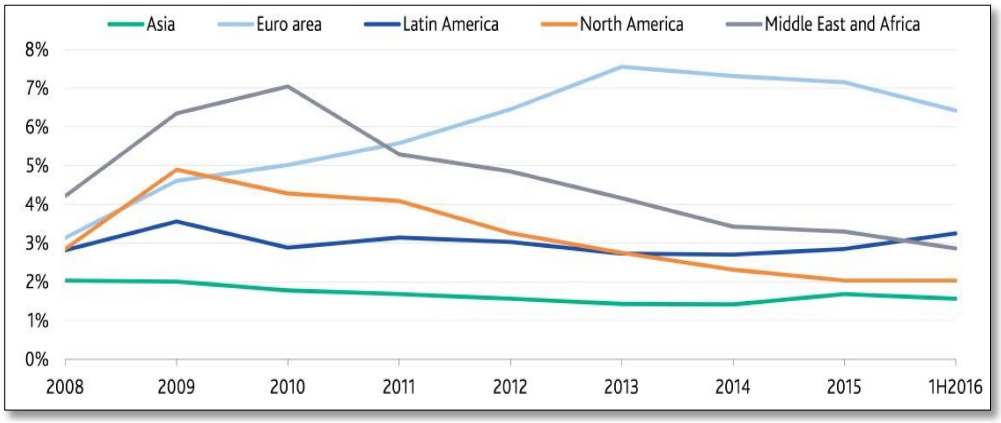


Figura 1. Ratio de morosidad del mundo. Fuente: Moddy's (2016).

Realidad Nacional

Según Kobsa (2017) especialista en el sector de cobranza, detectó que los créditos de las empresas en el Perú crecieron en 2.23% en total entre julio del 2016 y del 2017 sin embargo, los niveles de deuda siguen aumentando.

El banco central de reserva (BCR) reveló que el nivel de endeudamiento de las familias peruanas equivale en promedio 2.3 veces sus ingresos, lo que estaría limitando la posibilidad de crecimiento del consumo, ya que la situación económica de las personas no es estable. El principal riesgo es el sobreendeudamiento de los agentes económicos, de algunas familias y empresas que pueden sobre endeudarse más de sus posibilidades.

ASBANC (2017) la institución gremial que agrupa a los principales bancos e instituciones financieras privadas del Perú, señaló que la tasa de morosidad bancaria creció 2.96% al cierre de enero del 2017 lo que significó un incremento del 0.16% por encima de la tasa de diciembre del año pasado. El resultado habría sido explicado por un incremento en la morosidad de los créditos de las grandes empresas un 1.09%, medianas empresas un 6.74% y pequeñas empresas 9.04% y de los préstamos a personas, como créditos de consumo un 3.74% y los hipotecarios un 2.32%.

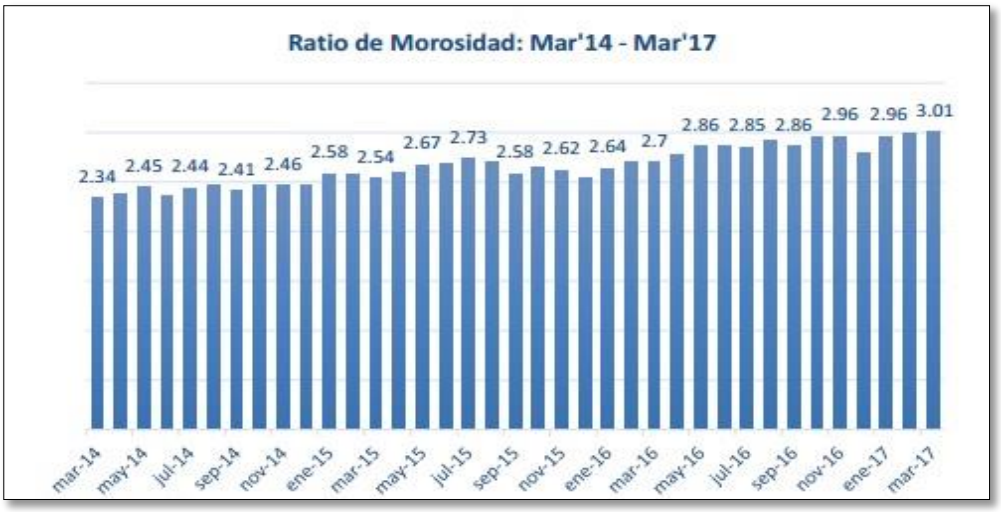


Figura 2. Ratio de morosidad en el Perú. Fuente: ASBANC (2017).

Realidad Local

Oncosalud es la primera empresa oncológica del Perú especializada en prevención, diagnóstico y tratamiento del cáncer, así como también el desarrollo de programas oncológicos de tal manera que las personas puedan gozar de la mejor atención. Entre los servicios que brindan son: prevención del cáncer, detección a tiempo, diagnóstico, tratamiento eficientemente y cuidado del paciente en su recuperación.

La empresa Oncosalud se encarga de brindar servicios médicos relacionados con enfermedades oncológicas para todo tipo de personas, pero al vender estos seguros a los clientes se presentan deficiencias al momento del pago de la misma ya que sobrepasan la fecha indicada, desatando altos niveles de morosidad de los clientes.

Al realizar un análisis se detectó que la empresa de seguros Oncosalud tiene un alto índice de incumplimiento de pago por parte de los clientes que se encuentran afiliados a la empresa. Según el departamento de cobranza de la empresa Oncosalud el 35% de sus clientes tiene deuda en sus pagos ya sea con un tiempo de retraso de 1 mes, 2 meses o más, esto se debe a que la empresa no cuenta con una gestión preventiva que le ayude a conocer si un cliente puede ser moroso.

La presente investigación se realizará en la empresa de seguros Oncosalud, la cual está ubicada en la Av. Guardia Civil 571, San Borja.

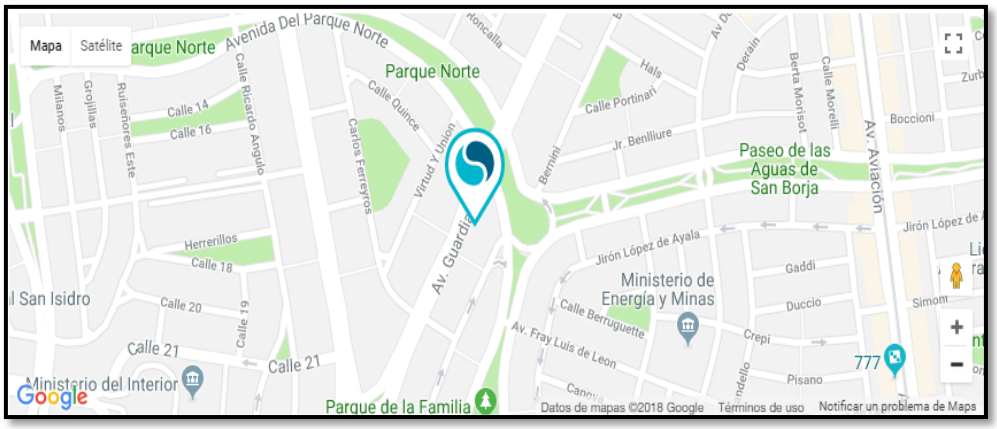


Figura 3. Ubicación de la empresa Oncosalud.

1.1.2. Definición del problema

La empresa de seguros Oncosalud cuenta con varias sucursales, tanto en Lima como en provincia. El departamento de cobranza de la empresa necesita conocer cierto tipo de información del cliente con respecto a la morosidad de cada uno de ellos. Debido a que la empresa no cuenta con una herramienta que le ayude a conocer que clientes podrían ser morosos, esto ha llevado a una alta cifra de morosidad.

Su problemática se basa en no tener información importante sobre los clientes, como para saber si en algún momento le será perjudicial. La empresa necesita dicha información para analizar dicho resultado y luego tomar mejores decisiones en base a ello.

La mayor cantidad de sus ventas son por medio de empresas terciarias (call centers) y no se preocupan a que clientes les están vendiendo los seguros, es decir no analizan la información del cliente, por este motivo la empresa hoy en día presenta cifras de clientes morosos.

Como se ve en la figura 4, la empresa de seguros Oncosalud tiene una alta cifra de morosidad en los primeros cuatro meses del 2018, esto

debido a que la empresa no cuenta con herramientas que le ayuden a predecir que clientes serán más propensos a caer en morosidad.

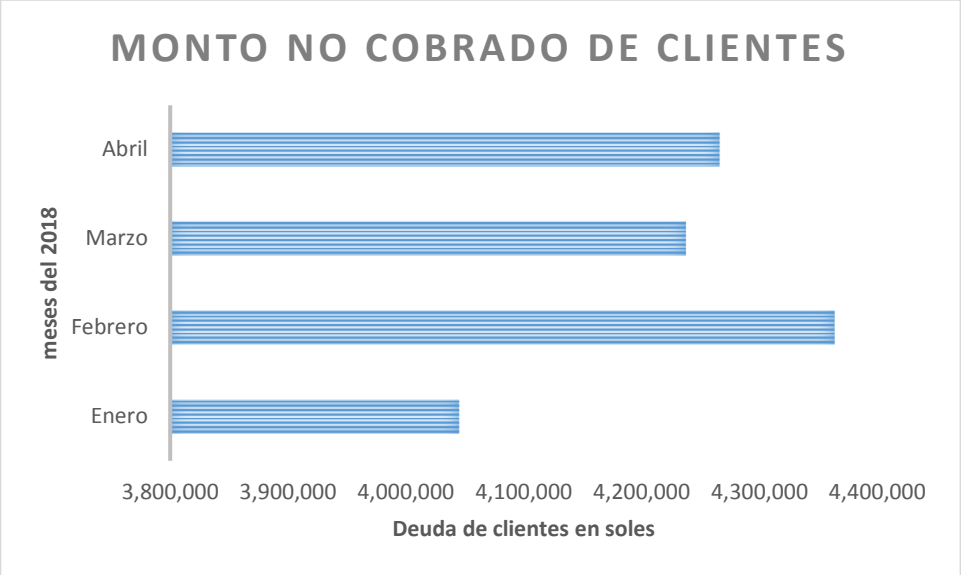


Figura 4. Monto no cobrado de clientes de la empresa Oncosalud.

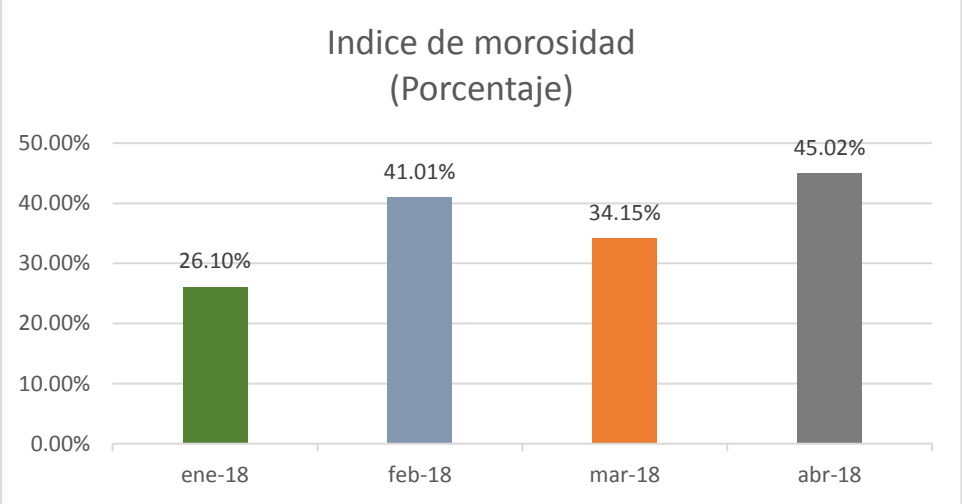


Figura 5. Índice de morosidad entre los meses enero y abril 2018.

En la figura 5 muestra el índice de morosidad de la empresa, están evaluados mediante el promedio de rango de mora inicial de los clientes en el cual se visualiza que en el mes de abril de 2018 han tenido un alto porcentaje de índice de morosidad. Por esta razón, se utilizaran las técnicas de la minería de datos para predecir el riesgo de morosidad de los clientes de la empresa y tomar mejores decisiones que se requieran.

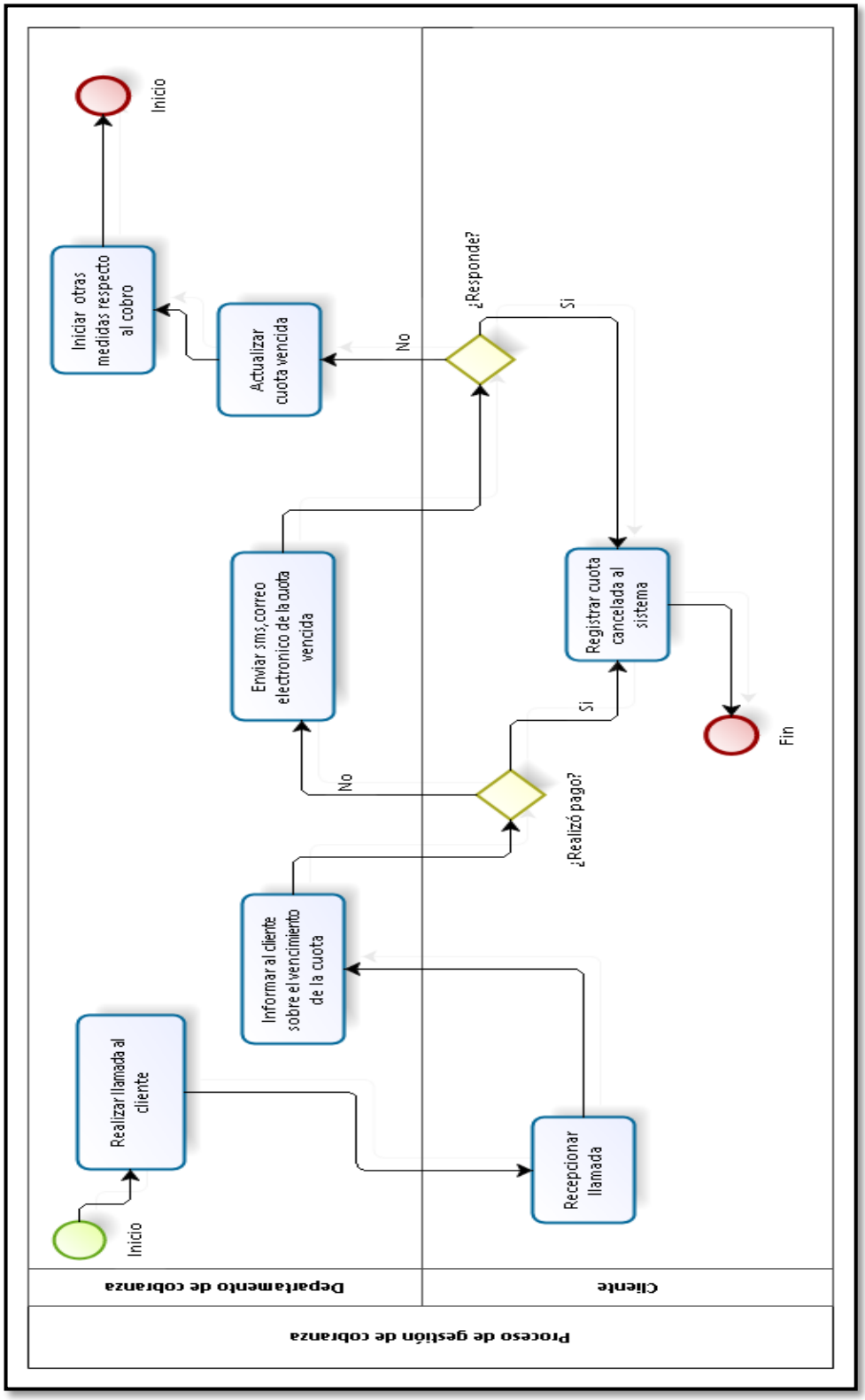


Figura 6. Proceso de cobranza de la empresa Oncosalud.

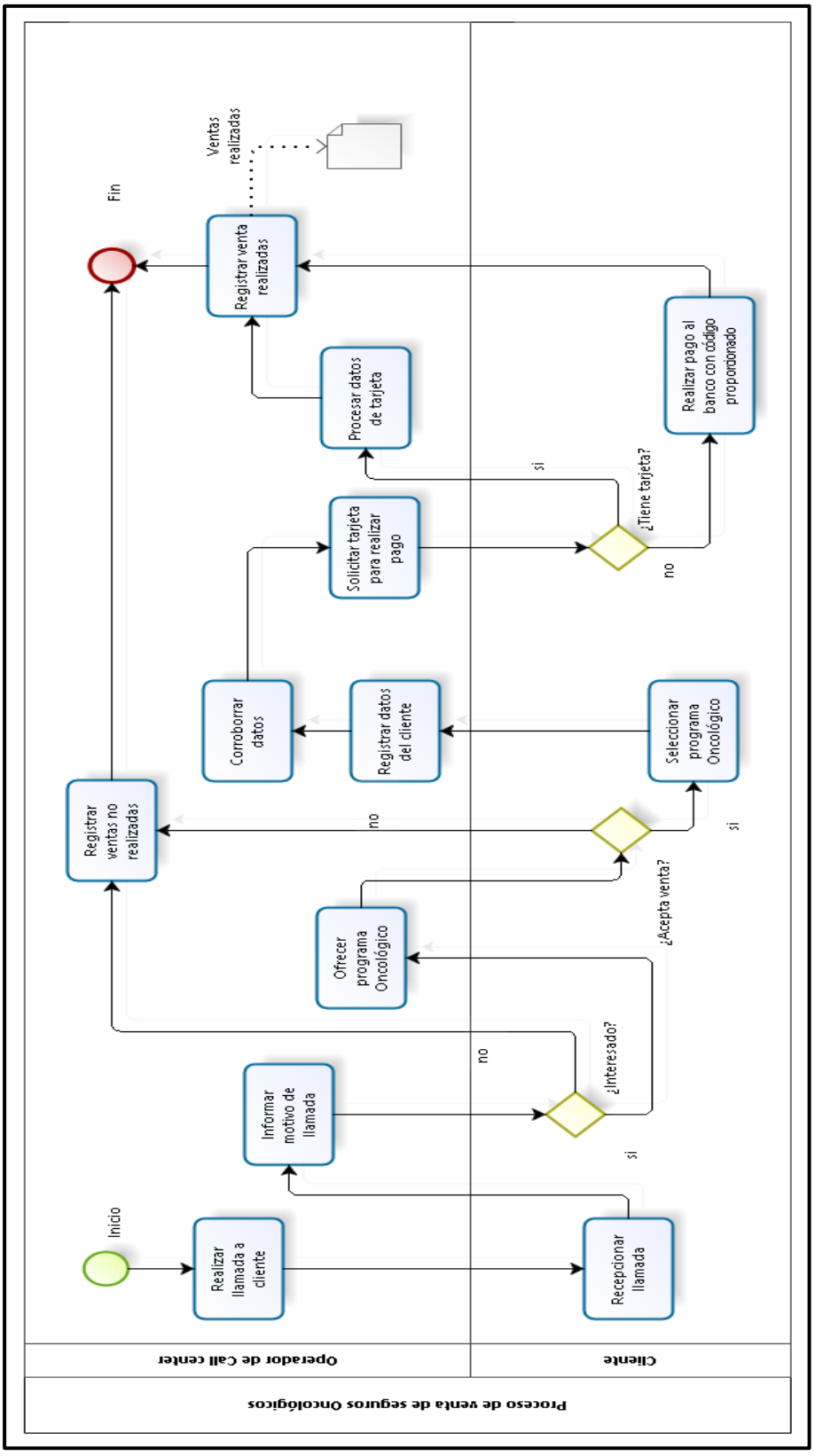


Figura 7. Proceso de ventas de seguro Oncológicos.

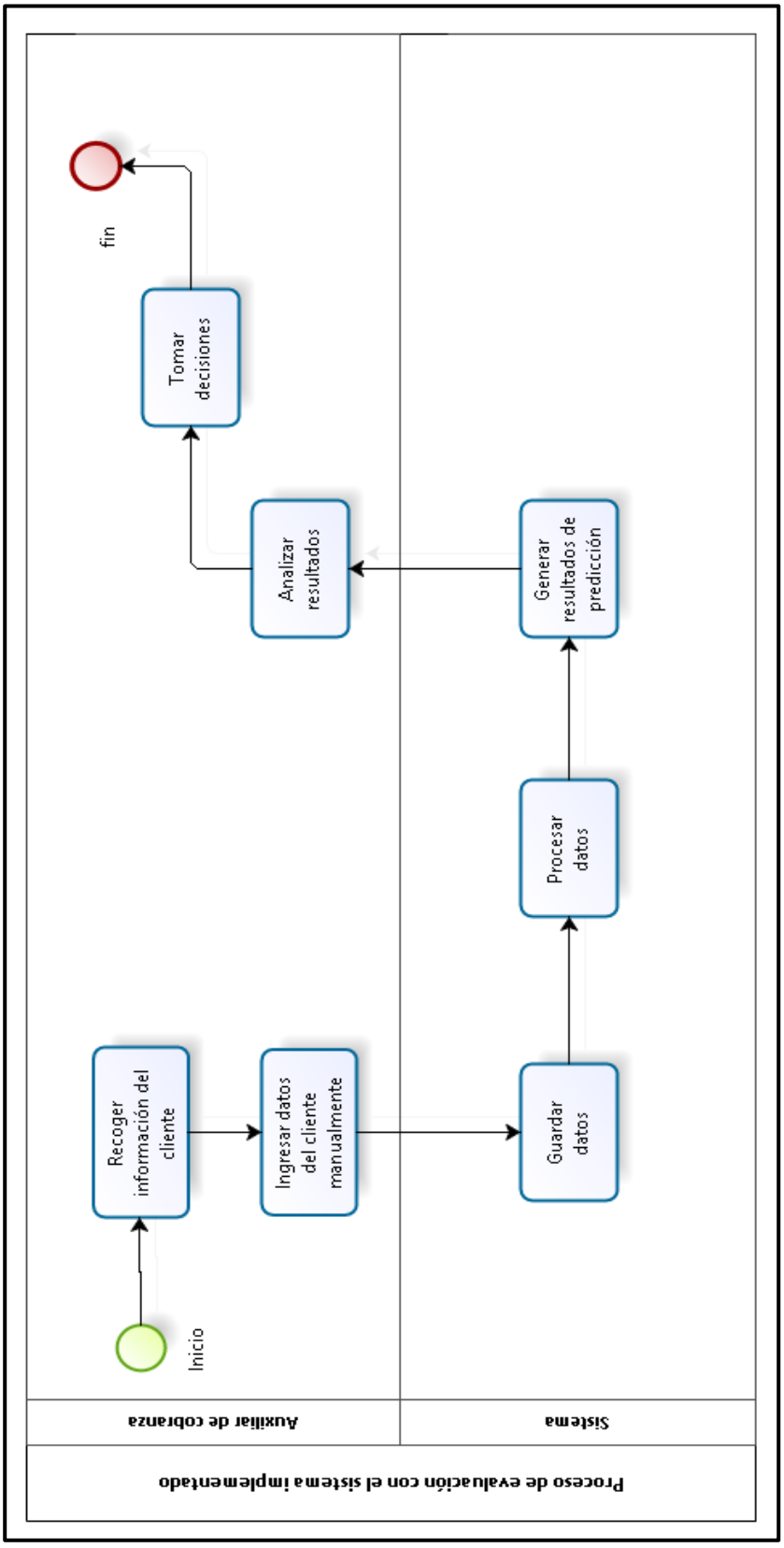


Figura 8. Proceso de evaluación con el sistema implementado.

1.1.3. Enunciado del problema

¿En qué medida la aplicación de minería de datos basado en árboles de decisión facilitará la predicción del riesgo de morosidad de los clientes en la empresa de seguros Oncosalud SAC 2018?

1.2. TIPO Y NIVEL DE LA INVESTIGACIÓN

1.2.1. Tipo de Investigación

Aplicada:

De acuerdo al propósito de la investigación, el problema y los objetivos formulados, reunió las condiciones para tratarse como una investigación aplicada, ya que se desarrolló y se apoyó en la parte teórica es decir desarrollar un modelo de minería de datos basado en árboles de decisión con la finalidad de dar una solución factible al problema de la predicción del riesgo de morosidad.

1.2.2. Nivel de Investigación

Explicativo:

En el presente trabajo busca explicar cómo la variable independiente influye en la dependiente, dando un resultado favorable.

1.3. JUSTIFICACIÓN DE LA INVESTIGACIÓN

El impacto del impago de los clientes es uno de los grandes desafíos a los que deben enfrentarse las organizaciones y que puede resultar un problema muy grave hoy en día. Los problemas de endeudamiento de los clientes afectan directamente la situación financiera y que pueden causar hasta el cierre de una organización.

En la empresa Oncosalud los clientes no están cumpliendo con sus obligaciones de pago y cancelan sus deudas después de la fecha de vencimiento establecido, por lo tanto, en el presente trabajo se aplicará la técnica minería de datos basado en árboles de decisión para dar solución

a los problemas de morosidad que aquejan a la organización, ya que no cuentan con herramientas tecnológicas que le ayuden a predecir el cuál amerita hacerlo más automatizado, y eficiente aplicando minería de datos. La técnica aplicada en el desarrollo de este trabajo permitirá prospectar a los clientes para así poder determinar que clientes serán perjudiciales a la empresa, evitando problemas de endeudamiento que pueden afectar en un futuro a la entidad.

Por último, la investigación del proyecto ha sido elegida con el fin de tomar mejores decisiones y evitar la pérdida de clientes en la empresa, estas decisiones serán tomadas una vez aplicada la técnica de minería de datos basado en árboles de decisión utilizando la información histórica de los clientes, que harán mitigar posibles riesgos de morosidad y a su vez mejorar el rendimiento financiero de la organización, es entonces que debemos comprender que en estos tiempos es de suma importancia contar con las herramientas necesarias para la mejora de su entidad.

1.4. OBJETIVOS DE LA INVESTIGACIÓN

1.4.1. Objetivo general

Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión facilita la predicción del riesgo de morosidad de los clientes en la empresa de seguros Oncosalud SAC 2018.

1.4.2. Objetivos específicos

- Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión mejora la precisión de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud.
- Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión reduce la dificultad de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud.
- Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión reduce el tiempo de predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

1.5. HIPÓTESIS

La aplicación de minería de datos basado en árboles de decisión facilita significativamente la predicción del riesgo de morosidad de los clientes en la empresa de seguros Oncosalud SAC 2018.

1.6. VARIABLES E INDICADORES

1.6.1. Variable Independiente

a. INDICADORES

Variable Independiente: Árboles de decisión

Tabla 1

Conceptualización de la variable Independiente

Indicador: Presencia – Ausencia
Descripción: Cuando se indique No, significa que no ha sido desarrollado la aplicación de árboles de decisión para predecir el riesgo de morosidad y todavía el problema se encuentra en la situación actual. Cuando indique Sí, es cuando se ha desarrollado el modelo para predecir el riesgo de morosidad de clientes.

b. ÍNDICE

Variable Independiente: Árboles de decisión

Tabla 2

Indicador Variable Independiente

INDICADOR	INDICE
Presencia – Ausencia	No, Si

1.6.2. Variable Dependiente

a. INDICADORES

Variable dependiente: Predicción del riesgo de morosidad

Tabla 3

Conceptualización de indicadores de la variable dependiente

Indicadores	Conceptualización
Precisión de predicción	Mide el éxito de una experiencia respecto a las predicciones realizadas.
Error de predicción	Este indicador mide la fracción de datos mal clasificados en datos objetivos.
Nivel de dificultad	Es el grado de dificultad o inconveniente que se presenta para realizar acciones y cumplir un propósito.
Tiempo para predecir	Es el tiempo que se requiere para estimar el riesgo de morosidad de un cliente.

b. ÍNDICE

Variable dependiente: Predicción del riesgo de morosidad

Tabla 4

Indicador Variable Dependiente

Indicador	Índice	Unidad de Medida	Unidad de observación
Precisión de predicción	[60% - 95%]	Porcentaje	Reportes, Software
Error de predicción	[1% - 40%]	Porcentaje	Reportes, Software
Nivel de dificultad	[Muy difícil, Difícil, Normal, Fácil, Muy fácil]	Escala de Likert	Cuestionario
Tiempo para predecir	[1 - 780]	Segundos	Cronometro

1.7. LIMITACIONES DE LA INVESTIGACIÓN

Para la realización del presente trabajo de investigación se presentaron las siguientes limitaciones, las cuales fundamentalmente se reúnen en:

- La base de datos adquirida de la empresa de seguros Oncosalud es a nivel nacional, dado que para el estudio solo se cuenta con la base de datos de las sedes de lima.
- La empresa de seguros Oncosalud solo brinda información parcial motivo al uso de su reglamento interno por parte del MOF (Manual de organización y funciones).

1.8. DISEÑO DE LA INVESTIGACIÓN

Pre-Experimental: Porque se demostrará la hipótesis utilizando un solo grupo mediante métodos experimentales.

Tabla 5

Diseño de la investigación

Ge	O1	X	O2
Clientes de la empresa de Seguros Oncosalud	Pre-prueba o medición previa al estímulo o tratamiento experimental	Riesgo de morosidad	Post-prueba o medición posterior al estímulo o tratamiento experimental

Donde:

Ge: Grupo experimental: Es el grupo de estudio al que se aplicará el estímulo (Técnica de minería de datos).

O1: Son los valores de los indicadores de la variable dependiente en la pre-prueba.

X: La implementación de la técnica de minería de datos: Estimulo o condición experimental.

O2: Son los valores de los indicadores de la variable dependiente en la post-prueba (después de haber implementado la técnica de minería de datos).

Descripción:

Consiste en la comparación de un grupo experimental (Ge) constituido por un número representativo de clientes del mes de julio de la empresa de seguros Oncosalud, cuyos indicadores se les realiza una Pre – Prueba (O1). Después se le aplicará un estímulo de la aplicación de árboles de decisión (X); finalmente se aplicará una nueva medición de los indicadores (O2).

1.9. TÉCNICAS E INSTRUMENTO PARA LA RECOLECCIÓN DE INFORMACIÓN

1.9.1. Técnicas e Instrumentos

Tabla 6

Técnicas e instrumentos de la investigación

Técnicas	A quien se aplica	Instrumentos	Método	Indicador medido
Revisión de documentos	Proceso de recupero de clientes, Software	-Reportes	1. Seleccionar los documentos a revisar 2. Registrar los datos relevantes 3. Analiza la información 4. Ejecutar las formulas	-Precisión de predicción -Error de predicción
Aplicación de encuestas Preguntas Cerradas	Empleados del departamento de cobranza	Cuestionario **APÉNDICE I**	1. Elaborar el cuestionario 2. Validar el cuestionario con expertos 3. Ejecutar la encuesta 4. Analizar resultados	-Nivel de dificultad
Observación directa con cronometro	Proceso de recupero de clientes, Software	Ficha de registro **APÉNDICE II**	1. Ubicar la zona de estudio más apropiada 2. Definir objeto de observación 3. Elegir instrumentos de observación 4. Analizar datos	-Tiempo para predecir

CAPÍTULO II
MARCO TEÓRICO

2.1. ANTECEDENTES DE ESTUDIOS

Los proyectos que se han investigado anteriormente encontramos autores que brindan información con temas referentes a nuestra investigación lo cual da un gran aporte al trabajo que estamos desarrollando.

A continuación, presentaremos las tesis de investigación que hace referencia a nuestro proyecto:

Antecedentes Internacionales

A. Autor: Córdova Galarza Janeth Carolina

Título: Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes que pertenecen al colegio fiscomisional "San Francisco" de la ciudad de Ibarra.

Correlación:

Ordoñez (2013) realizó un estudio sobre la deserción de los estudiantes y sus posibles factores que influyen en el alumno para poder tomar estas decisiones en el primer ciclo de la Universidad Técnica Particular de Loja. Esta investigación desarrolló un modelo de minería de datos basado en la metodología CRISP-DM, tomando como referencia la información de los diferentes estudiantes proporcionados desde la base de datos de su sistema académico, y a su vez obteniendo resultados importantes.

El objetivo de esta investigación fue obtener beneficios económicos y determinar estrategias con respecto a la deserción de los estudiantes. La muestra que se tomó en esta investigación fueron las carreras que poseen mayor número de estudiantes, cursando cinco cursos y cumpliendo la modalidad abierta y a distancia.

Lo resultados que se obtuvieron en la presente investigación indicaron que los estudiantes que poseen la edad de 16 y 26 años son los que desertan con mayor frecuencia, a su vez también se pudo determinar que los estudiantes que inician su carrera a menor edad mayor es la probabilidad que deserten ya que no tuvieron la orientación adecuada para empezar una carrera universitaria. Otro de los resultados de la

investigación fueron que el 81.51% del área técnica presenta mayor deserción debido a que esa posee un alto nivel de esfuerzo y dedicación, ya que dicha área posee complejidad de análisis y entendimiento.

En conclusión, el antecedente presenta alguna proximidad con el presente estudio en el sentido que abarca el desarrollo de un modelo de minería de datos, razón por la cual se guió en la elaboración del marco teórico referida a la variable independiente.

B. Autor: Díaz Avendaño, Ángel Arnulfo

Titulo: Técnicas de Minería de datos para predicción del diagnóstico de hipertensión arterial.

Correlación:

Díaz (2016) realizó una investigación sobre la hipertensión arterial, una de las importantes causas de muertes en el mundo. Según el reporte de la organización Mundial de la Salud del 2012, señala que 1 de cada 3 personas en el mundo sufren de hipertensión arterial. La hipertensión arterial es la segunda causa de muerte a nivel mundial y considerado como “muerte silenciosa”.

En la investigación desarrollada se aplicó la metodología KDD para la extracción Automática de Conocimiento y más concretamente se centra en la etapa de Minería de Datos (MD). El objetivo de la investigación fue encontrar patrones en la información, con el fin de crear modelos en los cuales se basó en las reglas de asociación y árbol de decisión.

Como muestra para dicha investigación se recopiló información de los pacientes del centro hospitalario “Almanzor Aguinaga Asenjo – Chiclayo”, brindando una gran cantidad de datos con un total de 8,735 registros para su proceso de estudio, las técnicas que se utilizaron para la investigación fueron árboles de decisión y asociación.

Como resultado de la investigación del diagnóstico proyectado se concluyó que 749 pacientes del centro hospitalario estuvieron más propensos a padecer de hipertensión arterial y de la representación se

obtuvo un nivel de confianza del 97% con las técnicas de árboles de decisión y 760 pacientes con un 98.6% con las técnicas de asociación.

En conclusión, el antecedente investigado presentó una proximidad con el estudio desarrollado en el sentido que proporcionó información importante acerca de su metodología planteada en el desarrollo de su investigación, por lo tanto, se tomó como guía para el estudio que se viene abordando.

C. Autor: Ordoñez Briceño Karla Fernanda

Título: Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL - ECUADOR.

Correlación:

Ordoñez (2013) realizó un estudio sobre la deserción de los estudiantes y sus posibles factores que influyen en el alumno para poder tomar estas decisiones en el primer ciclo de la Universidad Técnica Particular de Loja. Esta investigación desarrolló un modelo de minería de datos basado en la metodología CRISP-DM, tomando como referencia la información de los diferentes estudiantes proporcionados desde la base de datos de su sistema académico, y a su vez obteniendo resultados importantes.

El objetivo de esta investigación fue obtener beneficios económicos y determinar estrategias con respecto a la deserción de los estudiantes. La muestra que se tomó en esta investigación fueron las carreras que poseen mayor número de estudiantes, cursando cinco cursos y cumpliendo la modalidad abierta y a distancia.

Lo resultados que se obtuvieron en la presente investigación indicaron que los estudiantes que poseen la edad de 16 y 26 años son los que desertan con mayor frecuencia, a su vez también se pudo determinar que los estudiantes que inician su carrera a menor edad mayor es la probabilidad que deserten ya que no tuvieron la orientación adecuada para empezar una carrera universitaria. Otro de los resultados de la

investigación fueron que el 81.51% del área técnica presenta mayor deserción debido a que esa posee un alto nivel de esfuerzo y dedicación, ya que dicha área posee complejidad de análisis y entendimiento.

En conclusión, el antecedente presenta alguna proximidad con el presente estudio en el sentido que abarca el desarrollo de un modelo de minería de datos, razón por la cual brinda un gran aporte en la elaboración del marco teórico referida a la variable independiente e interviniente.

Antecedentes Nacionales

D. Autor: Irene leydi, Roque Montalvo

Título: Análisis comparativo de técnicas de minería de datos para la predicción de ventas.

Correlación:

Roque (2016) realizó una investigación sobre la comparación entre distintas técnicas utilizadas en la minería de datos para el diseño de modelo de pronósticos de series de tiempo. El ámbito de este estudio se centra en la empresa “El Astro S.A.C.” para determinar las estimaciones de ventas según el volumen que genera mensual o trimestral. En la actualidad existen diferentes técnicas para la generación de pronósticos de series de tiempo, desde los modelos de tipo estadísticos, o los más avanzados que usan algoritmos computacionales basados en inteligencia artificial como las redes neuronales o las máquinas de soporte vectorial. El problema no trata sobre la construcción de un modelo de minería de datos, si no de evaluar que algoritmo y técnica sirve o tiene un mejor performance para un problema determinado.

En esta investigación se habló sobre dos metodologías que es la metodología KDD (comprendida por sus 5 procesos) y la metodología CRISP-DM (comprendida por sus 6 procesos), y a su vez se evaluaron algoritmos para poder comparar cuál de ellas tiene una mejor performance para un problema determinado.

Esta investigación tenía como objetivo realizar un análisis comparativo acerca del rendimiento de las técnicas para la predicción de ventas a la comercialización de artículos deportivos, en esta investigación no se requirió muestreo.

Como resultado de la investigación se realizó la evaluación de las técnicas de minería de datos Regresión, Series temporales, Redes Neuronales, Agrupamiento teniendo como resultado que las técnicas más adecuadas para el ámbito de ventas orientado a la comercialización de artículos deportivos en esta investigación son las de serie de tiempo.

En conclusión, este antecedente específico presenta algunos fragmentos importantes para el desarrollo de la investigación, razón por la cual se ajusta a la construcción de estudio referida a la variable interviniente.

E. Autor: María Gabriela Camborda Zamudio

Título: Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la carrera de ingeniería civil de la universidad continental.

Correlación:

Camborda (2014) realizó una investigación acerca del rendimiento académico de los estudiantes de la universidad continental que cursan los primeros ciclos de la carrera de ingeniería civil ubicada en la ciudad de Huancayo, de manera que para el desarrollo de esta investigación utilizó información en cuanto a los factores demográficos, académicos, institucionales y actitudinales de los estudiantes.

El desarrollo de esta investigación tuvo como objetivo aplicar árboles de decisión específicamente con el algoritmo J48 de weka para poder predecir los factores que más influyen en los estudiantes en cuanto a su rendimiento académico.

Como resultado de la investigación se obtuvo que los factores académicos poseen mayor ganancia de información y mayor exactitud

en su predicción, la cual esta variable presenta mayor relevancia ya que ayuda a definir el rendimiento académico del estudiante.

En conclusión, el antecedente presentó alguna proximidad con esta investigación, de manera que aporta como guía en el desarrollo del marco teórico en cuanto a la variable independiente “árboles de decisión” de la técnica de minería de datos.

F. Autor: Chero Vásquez, Keysi Berlyt y Paredes Abanto, María Elsa

Título: Estrategias crediticias para disminuir el índice de morosidad en el banco azteca, Chepén 2015.

Correlación:

Vásquez y Paredes (2015) realizaron una investigación acerca de la morosidad de los clientes que se daba en la institución financiera del Banco Azteca, Chepén 2015. Esta investigación tuvo como finalidad minimizar el índice de morosidad de los clientes mediante métodos financieros.

La muestra que se utilizó para el desarrollo de esta investigación fue conformada por colaboradores del banco azteca de Chepén en el área de cobranza. Como herramientas de captación de datos se utilizó listas de cotejo para las variables dependientes con el objetivo de conocer las dificultades y de qué modo pueden ser optimizadas y así minimizar la morosidad, donde se aplicaron diferentes estrategias.

Los resultados que se obtuvieron mostraron que la aplicación de los métodos financieros propuestos alcanzó aumentar del 67% a 80%. Por lo tanto, se concluyó que el índice de morosidad ha reducido de un 60% a 23%, en ese sentido se observó que la investigación realizada permitió incrementar el retorno de los créditos otorgados.

En conclusión, el antecedente investigado muestra relevancia y aporte al estudio que se está desarrollando, razón por la cual tomamos como guía a esta investigación ya que se ajusta a la variable dependiente que

es “predicción del riesgo de morosidad” para la elaboración del marco teórico.

G. Autor: Julio Cesar Carpio Ticona

Título: Modelo de predicción de la morosidad en el otorgamiento de crédito financiero aplicando la metodología CRISP – DM.

Correlación:

Carpio (2016) realizó una investigación acerca del impago de los clientes que pueden producir crisis o hasta el quiebre de las instituciones financieras. En este estudio se desarrolló un modelo de predicción aplicando la técnica de minería de datos, la cual emplearon información auténtica de los clientes, identificando las variables principales para el desarrollo de su investigación.

La metodología que se utilizó para dicha investigación fue CRISP-DM conformada con sus seis fases para la guía metodológica del desarrollo y selección del mejor algoritmo que se utilizaron en este estudio.

El objetivo de la investigación que se realizó es tomar futuras decisiones respecto al otorgamiento de crédito financiero en la institución Caja Rural de Ahorro y Crédito Los Andes (CRAC Los andes).

La muestra que se utilizó para el desarrollo de la investigación se usó la población total (846 registros), razón por la cual la muestra no ha sido calculada sin embargo para la validación cruzada del modelo se utilizó métodos de grupo y para ello se utilizó grupos de 10 de la población.

Como resultado en esta investigación se logró cuatro algoritmos computacionales, dichos algoritmos presentaron un grado de certeza por encima del 78% en la predicción de crédito, el algoritmo “Bosques aleatorios” presentó un grado de certeza global del 82% mediante la matriz de confusión.

En conclusión, esta investigación presenta similitud con el estudio que se viene abordando en el sentido que se ajusta a la variable independiente y dependiente que son las técnicas de minería de datos y

riesgo de morosidad, asimismo guio en la construcción de la metodología de estudio.

2.2. DESARROLLO DE LA TEMÁTICA CORRESPONDIENTE AL TEMA INVESTIGADO

2.2.1. Riesgo de morosidad

2.2.1.1. Morosidad

La morosidad desde el punto de vista empresarial, es el incumplimiento y el atraso que tiene un individuo ante sus obligaciones de pagos financieros en la fecha pactada contratada. El riesgo de crédito deriva del índice de morosidad, el cual tiene por significado como la medida de la cartera vencida con relación a la cartera total de la empresa.

El desarrollo del índice de morosidad se debe a la forma de trabajo, políticas, objetivos y recursos de cada entidad. Según Chavarrín (2015) menciona que los factores importantes son los económicos, políticos y regulatorios.

Según Chavarrín (2015) emplea: “Como indicador del riesgo de crédito al índice de morosidad” (p.76). Ya que es una variable negativa y significativa. En base a ello las entidades son muy limitantes al dar un crédito, con la finalidad de que no crezca el índice de morosidad y no salga perjudicada la entidad.

2.2.1.2. Características de la morosidad

Según Kabari (2009) clasifica a las personas que solicitan un préstamo en buenos y malos candidatos. Los buenos son las personas que piden un préstamo y tienen una probabilidad alta a que devuelvan el préstamo en el tiempo estipulado. Mientras que los malos candidatos deben ser rechazados ya que no devolverían el préstamo en el tiempo solicitado.

Generalmente las entidades otorgan crédito en base a su simple conocimiento, ya sea utilizando historiales anteriores, que le ayuden a predecir si el cliente podría fallarles en algún momento, en base a ello se

toman las decisiones pertinentes en la entidad. Sin embargo, hay algunas desventajas al realizar este procedimiento como menciona Kabari (2009) alto costo de oficiales de crédito de formación, decisiones incorrectas, largo periodo de tiempo para evaluar la categoría de riesgo del cliente y tomar la decisión de concesión de crédito.

2.2.1.3. Causas de la morosidad

Como argumenta el morosólogo español Brachfield (2014) quien es el creador de la morosología, a su vez es uno de los mayores especialistas contra la morosidad en España menciona que las causas principales por las que existen morosos son:

- **Causa de iliquidez y problemas financieros:**

  El deudor no dispone de efectivo suficiente e inmediato para cumplir con un pago al momento de su vencimiento ya sea pagos operativos o financieros. Si el deudor se mantiene en esta situación se encontrará ante un problema económico. Es por eso que, si el deudor no tiene medios para pagar sus deudas, se convertirá en moroso ya que tendrá un retraso en sus pagos y este problema se vuelve más crónico. Por eso el deudor tendrá que aumentar sus ingresos para superar esa situación.

- **Causas económicas:**

  Esta causa ya es un problema más grave que la situación financiera que afecta seriamente al deudor. Los ingresos o bienes que tiene el deudor no son suficiente para cubrir las deudas de este, es por ello que afecta en sus pagos, ya que no cuenta con efectivo al momento del vencimiento del pago.

- **Causas circunstanciales:**

  El deudor tiene problemas que suceden de un momento a otro, son problemas que no se tienen previstos, el cual se necesita de un gasto extra para poder solventar estas dificultades, es por esos motivos que se tiene un retraso en los pagos. Algunas dificultades

coyunturales son: una enfermedad, una pérdida, una mala compra, etc.

- **Causas culturales:**

  El deudor tiene los medios para poder cancelar sus deudas, pero por una cultura incorrecta no se le da por estar al día con la empresa. Ya que el deudor tiene en mente esta cultura piensa que dicha actitud es normal.

- **Causas de nivel intelectual:**

  Algunos clientes o deudores no cuentan con información necesaria y el conocimiento para entender lo que implica que no se cumpla con las fechas establecidas del pago y que estas pueden afectar a la entidad.

- **Causas emocionales:**

  Muchos deudores disponen del capital para pagar sus deudas, pero no lo cumplen pese a distintos motivos como el malestar con la entidad, la desconfianza e interacciones que resultaron tediosas. Estas causas se ocasionan porque el cliente considera que los pagos o deudas son injustos y piensan que realizando eso será un castigo para la entidad. También puede ocurrir que el deudor no considere justa la deuda y por eso no pague la deuda.

2.2.1.4. Tipos de morosos

Según Brachfield (2014) director de estudios de la plataforma multisectorial contra la morosidad (PMcM) clasifica a los morosos en cinco tipos:

- **Moroso intencional:** Es la persona que puede pagar, pero no quiere.
- **Moroso fortuito:** Es la persona que quiere pagar, pero no tiene dinero para hacerlo. Son las personas que cuando tienen medios económicos pagan la deuda.

- **Moroso incompetente y/o desorganizado:** Es la persona que puede pagar, pero no saben lo que tienen que pagar son muy desorganizados y despistados.

- **Moroso negligente:** Es la persona que no muestra interés en sus saldos pendientes y se endeudan, pero tienen la capacidad de poder pagar.

- **Moroso circunstancial:** Es la persona que puede pagar, pero no lo hace porque hay un litigio (disputa entre dos personas), han bloqueado el pago, pero pagaran cuando se solucione el problema con la entidad.

- **Moroso insumiso:** Es la persona que puede pagar, pero no quiere hacerlo ya que cree que no les corresponde pagar ya que la cantidad no es la correcta.

2.2.1.5. Clasificación por categoría de riesgo

Según Martínez (2013) clasifica el riesgo de crédito o insolvencia de clientes en función del riesgo de pérdida.

- **Normal o irregular:** Son aquellas operaciones que son impagados entre 1 y 90 días desde la fecha de vencimiento de la deuda. Su riesgo tiene una evidencia objetiva y verificable (Martínez, 2013).
- **Subestándar:** Son aquellas operaciones que presentan debilidades que pueden suponer la elevación de pérdidas de la entidad financiera mayor al nivel de protección que tienen las entidades.
- **Dudoso:** Son aquellas que son impagadas entre 90 días y 4 años, presentan desgaste de la solvencia del deudor y dudas razonables sobre su reembolso del principal e interés y comisiones.
- **Muy dudoso o Fallido:** Son aquellos que son impagados más de 4 años, después de analizarlo se dan de baja del balance de los activos de la entidad.

2.2.1.6. Dimensiones de la predicción del riesgo de morosidad

Las dimensiones trabajadas en esta investigación han sido revisadas en trabajos previos de los autores Díaz (2016), Gamarra (2014) y Roque (2016) cada uno con diferentes tesis, donde han utilizado las dimensiones que se están utilizando en la presente investigación.

1. **Precisión:** Es el grado de coincidencia existente entre los resultados independientes de una medición.

   - **Precisión de predicción:** Mide el éxito de una experiencia respecto a las predicciones realizadas, se calcula mediante la fórmula 1.

     Como se mide:

     $$Precisión = \frac{TP}{TP+FP} \quad (1)$$

     Donde:

     TP: Numero de valores correctamente predichos.

     FP: Numero de valores incorrectamente predichos.

   - **Error de predicción:** Este indicador mide la fracción de datos mal clasificados en datos objetivos, se calcula mediante la fórmula 2.

     $$eprecisión = 1 - precisión \quad (2)$$

2. **Dificultad:** Según Pérez y Merino (2008) define como problema que: “Surge cuando una persona intenta lograr algo, por lo tanto, son inconvenientes o barreras que hay que superar para conseguir un determinado objetivo”.

   - **Nivel de dificultad:** Es el grado de dificultad o inconveniente que se presenta para realizar acciones y cumplir un propósito.

3. **Tiempo:** Según Vicente define: “El tiempo es la magnitud física que permite secuenciar hechos y determinar momentos y cuya unidad de medida es el segundo”.

- **Tiempo para predecir:** Es el tiempo que se requiere para estimar el riesgo de morosidad de un cliente.

2.2.2. Minería de Datos

La minería de datos (Data Mining, en inglés) es un proceso para el descubrimiento de patrones de grandes bases de datos, no son descubiertos plenamente en su primera etapa del proceso de la minería de datos, forma parte de un proceso mucho más amplio conocido como descubrimiento del conocimiento.

Muchos investigadores han realizado diversas definiciones con respecto a la minería de datos autores como:

Según Fayyad, Piatetsky y Padhraic (1996) definen la minería de datos como: “El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos”.

Según Paz, Ojeda, Badillo, Bonett y Heredia (2018) definen la minería de datos como: “Es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos” (p.2).

Según Marulanda, López y Mejía (2017) definen la minería de datos como: “El proceso de descubrir conocimiento útil y entendible, desde grandes bases de datos almacenados en distintos formatos, por medio de modelos inteligibles a partir de los datos” (p.5).

Hoy en día muchas organizaciones competitivas toman decisiones a ciegas al momento de realizar una fidelización con el cliente, estas decisiones a ciegas pueden traer grandes consecuencias económicas a diversas organizaciones, es entonces donde la minería de datos debe integrarse para poder batallar frente a estos problemas.

La minería de datos es una tecnología de manejo de análisis de información que aprovecha hoy en día el procesamiento,

almacenamiento, y transmisión de datos a grandes velocidades y a un bajo costo. Permite encontrar conocimiento en grandes volúmenes de datos y tomar buenas decisiones mejor fundamentadas (Altamiranda, Peña, Ospino, Volpe, Ortega y Cantillo, 2013).

2.2.2.1. Tareas y técnicas de la minería de datos.

Más allá de la información explotable directamente de los datos almacenados en las bases de datos, la minería de datos se encarga de derivar conocimiento relevante que originalmente se encuentra oculto, después de haber preparado los datos, explorarlos y analizarlos según sus diversas técnicas.

Es conveniente categorizar la minería de datos en: tareas predictivas y tareas descriptivas, cabe resaltar que esta categorización no es única.

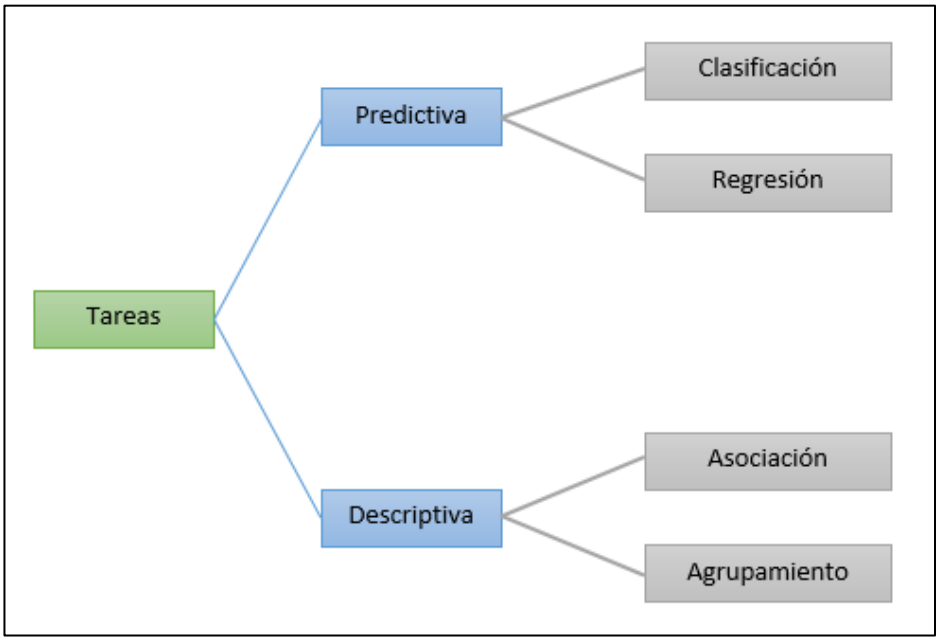


Figura 9. Tareas de la minería de datos.

- **Tareas supervisadas o predictivas.**

  Utilizan variables de una base de datos poder obtener estimaciones o pronósticos de comportamientos a futuros con respecto a los datos seleccionados, esta técnica resulta muy útil para poder predecir

valores desconocidos, por ejemplo, en aplicaciones para predecir el parte meteorológico o en la toma de decisiones por parte de un cliente en determinadas circunstancias (Aranda y Sotolongo, 2013).

Las tareas supervisadas se dividen en:

- **Técnica de clasificación**

  Definen unas series de clases, en que se pueden agrupar los diferentes casos. “Dentro de este grupo se encuentran las técnicas de árboles de decisión y reglas de inducción” (Aranda y Sotolongo, 2013, p.391). La técnica de clasificación es probablemente la tarea más familiar y más popular de la minería de datos donde son los encargados de encontrar modelos que distingan conceptos para futuras predicciones.

- **Técnica de regresión**

  Se usa una regresión para poder realizar una predicción de los valores ausentes de unas variables basándose en su relación con otras variables del conjunto de datos. La técnica de la regresión es parecida a la técnica de clasificación solo que la principal diferencia es que el valor que se va a predecir será numérico.

- **Tareas no supervisadas o descriptivas**

Las tareas descriptivas tienen como objetivo transformar el conjunto modelo (model set) en informaciones precisas, que reflejen las propiedades más relevantes y generalidades de los datos. “Estas tareas de MD, construyen modelos sobre patrones de hechos ocurridos en el pasado para su presentación de una forma comprensible” (Rivera, 2006). Buscan patrones humano-interpretables que describen los datos existentes.

Las tareas no supervisadas se dividen en:

- **Las técnicas de agrupamiento:** Consisten en agrupar datos dentro de un numero de clases que se realizan mediante criterios de distancia o similitudes, de manera que si las clases son semejantes entre sí se encuentres agrupadas (Jaramillo y Paz, 2015)**.** El objetivo de esta técnica es dividir los elementos en grupos basados en la semejanza, de tal manera que los grupos capturan la estructura natural de los datos.

- **Las técnicas de asociación:** En la minería de datos son empleadas en un conjunto de datos para hallar hechos que suceden en común mostrando posibles relaciones o correlaciones, es decir deben existir ciertas condiciones para que se origine dicha condición y también hallar reglas mediante un conjunto de datos que se generan por la naturaleza de las asociaciones y relaciones entre los datos de las entidades (Jaramillo y Paz, 2015).

2.2.2.2. Análisis predictivo

Según Espino (2017) refiere que el análisis predictivo consiste en extraer información de los datos, reutilizar estos datos para predecir patrones de comportamiento, aplicándose en cualquier evento desconocido, ya sea en el pasado, presente o futuro. Lo que realiza el análisis predictivo es identificar relaciones entre las diferentes variables de eventos y analizar dichas relaciones y predecir los resultados en situaciones futuras, para ello se debe analizar bien las variables ya que en base a ello saldrán los resultados de las predicciones.

El análisis predictivo genera conocimiento sobre las personas, ya que no solo trabaja con un solo dato sino con una gran cantidad de información y ayuda a identificar patrones de comportamiento. Un modelo predictivo es una herramienta que predice el comportamiento de un individuo. Según el autor mencionado el modelo predictivo “utiliza las características del individuo como entrada y proporciona una calificación predictiva como salida”.

2.2.2.3. Árboles de decisión

Según Barrientos et al. (2009) definen al árbol de decisión en un modelo de predicción el cual tiene por objetivo el aprendizaje inductivo, el cual se representa mediante un árbol, a partir de observaciones y construcciones lógicas. Es el modelo de clasificación más utilizado.

El resultado del uso del algoritmo diseña un árbol de decisión, por lo cual una de sus ventajas es la fácil interpretación y manera rápida de establecer si un cliente mediante una cierta cantidad de variables o datos históricos se encuentra bajo riesgo de caer en morosidad en la empresa (Contreras, Ferreira y Valle, 2017).

El árbol de decisión está conformado por un conjunto de nodos, hojas y ramas. El nodo principal o raíz del árbol es el atributo donde se inicia el proceso de clasificación. Los nodos internos vienen a ser las preguntas acerca del atributo en particular del problema. Es decir, las respuestas se representan mediante un nodo hijo. Las ramas entre cada nodo son los posibles valores del atributo. Los nodos hoja pertenecen a una decisión, la cual debe coincidir con una de las variables del problema a resolver.

Una vez que se ejecuta este tipo de modelo, solo habrá un camino el cual dependerá del valor actual de la variable estudiada.

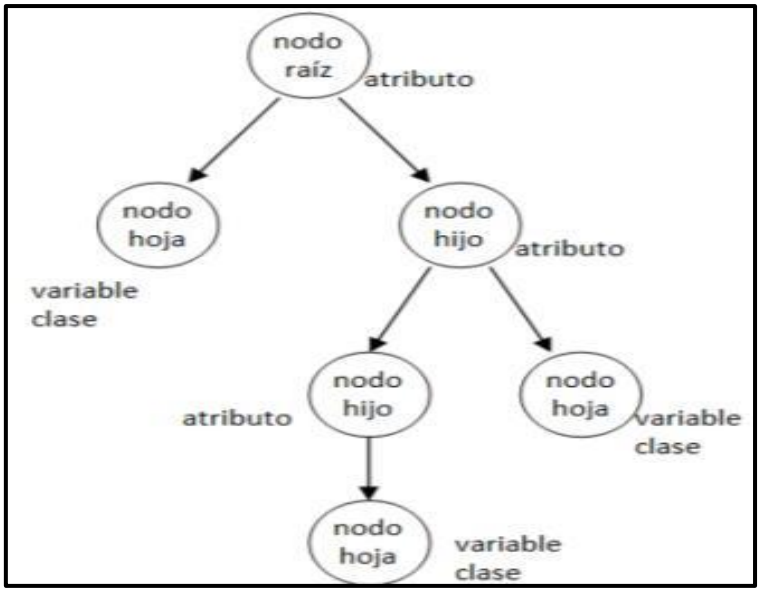


Figura 10. Estructura de un árbol de decisión. Vizcaíno (2008).

El algoritmo para generar los árboles de decisión tiene dos etapas:

- Inducción del árbol: En esta etapa se empieza a armar el árbol de decisión mediante un conjunto de preparación. Inicia generando el nodo raíz, eligiendo un atributo de prueba y dividiendo el conjunto de preparación en dos o más subconjuntos, para que se genere un nodo nuevo y así continuamente. Cuando un nodo tiene objetos de más de una clase se genera un nodo interno, cuando se tiene objetos de una clase se forma una hoja.

- Clasificación del árbol: En la siguiente etapa del algoritmo cada nuevo objeto es clasificado por el árbol construido, después se recorre todo el árbol (desde la raíz hasta una hoja). El camino que se debe tomar lo determina las decisiones tomadas en cada nodo interno, de acuerdo con el atributo de prueba presente.

Cómo funciona el algoritmo de árboles de decisión:

Para construir el algoritmo basado en árboles de decisión, se necesita la división de los datos. Una de las opciones para realizar dicha división se utiliza algunas medidas:

Entropía: Introducida por Shannon en su teoría de la información. El aspecto esencial de los algoritmos de árboles de decisión es escoger el mejor criterio al momento de dividir los datos, para esto se utiliza la entropía.

$$E(S)=\sum_{Ci=1}-pilog2(pi) \quad (3)$$

Donde:

S: conjunto de muestras

C: número de diferentes clasificaciones

$\mathbf{p_i}$: proporción de ejemplos que hay en la muestra

En el caso de una clasificación binaria es decir ejemplos positivos y negativos, la formula seria:

$$E(S) = -P\log_2(P) - N\log_2(N) \quad (4)$$

Donde:

P y N: número de ejemplos positivos y negativos

Luego se calcula la entropía de cada rama y se suman proporcionalmente las ramas para calcular la entropía del total:

$$E(T,X) = \sum c \in X p(c) E(Sc) \quad (5)$$

Ganancia de la información: Según Puncernau (2016) menciona: “Es la alternativa para determinar cuanta información se gana escogiendo el atributo como candidato a ser el nodo del árbol”.

Se resta este resultado a la entropía principal, obteniendo como resultado la ganancia de información usando dicho atributo.

$$\text{Gain}(T, X) = E(T) - E(T, X) \quad (6)$$

El atributo con mayor ganancia se selecciona como nodo principal o de decisión. Una rama con entropía 0 se convierte en una hoja. Es así como se construye poco a poco el árbol de decisión por completo.

Ventajas de árboles de decisión:

Algunas ventajas importantes que tienen los árboles de decisión:

- Los árboles de decisión son modelos de caja blanca: Una vez que se obtenga el modelo, es simple obtener de él una expresión matemática de fácil interpretación.
- Fácil interpretación, debido a la forma de mostrar el resultado.
- Permiten trabajar con pocos datos de entrenamiento: Puede dar un buen resultado a partir de pequeñas cantidades de datos.

- Poco costo computacional: Los algoritmos son bastantes eficientes y consumen pocos recursos de máquina.
- Herramienta integrable: Los árboles de decisión son una herramienta que se puede combinar fácilmente con otras herramientas de minería de datos.

2.2.2.4. Algoritmos para árboles de decisión:

Existen diversos algoritmos de árboles de decisión, los cuales se describen a continuación:

Decision Tree:

Según Beltrán y Poveda (2010) señala el árbol de decisión, es el método de clasificación con mayor potencial de uso, dado que es de fácil entendimiento. Para clasificar una serie de datos, el árbol realiza una revisión de la muestra desde los valores inferiores a los de mayor valor, cada nodo en el árbol de decisión es etiquetado con un atributo. De acuerdo al tipo de atributo, se determina el lugar jerárquico de cada nodo.

ID3:

El ID3 construye un árbol de decisión de arriba hacia abajo, sin hacer uso de backtracking (estrategia para encontrar soluciones) y se basa en ejemplos iniciales. Este algoritmo utiliza la variable ganancia de información para encontrar el atributo principal en cada paso. Es decir, tiene un método rápido para encontrar la mayor ganancia, la cual permite clasificar de manera adecuada.

Decision stump:

Este operador de aprendizaje, identifica aquellos arboles decisión con un solo nodo.

CHAID:

Según Berlanga, Rubio y Vila (2013) CHAID realiza un rápido árbol estadístico y multidireccional (varias direcciones), es decir utiliza datos de forma rápida y eficaz para crear perfiles del resultado esperado. CHAID elige la variable predictora la cual representará la interacción más fuerte. Las categorías de cada predictor no son utilizadas si no son importante para la variable dependiente.

J48 (C4.5):

El algoritmo J48 establecido en WEKA es una implementación del algoritmo C4.5, siendo uno de los algoritmos más utilizados en la minería de datos. El algoritmo C4.5 construye árboles de decisión usando el concepto de la entropía de la información (Tello, Eslava, y Tobías, 2013, p.20).

J48Graft:

El algoritmo J48Graft genera DT (Decision Tree) injertado de un árbol J48. Agrega nodos a un árbol de decisión existente con el propósito de reducir errores de predicción. Este algoritmo tiene la capacidad de identificar regiones que no están ocupadas, es decir realiza una nueva prueba en la hoja generando nuevas ramas que conducirán a nuevas clasificaciones, el injerto es un algoritmo para agregar nodos a un árbol como un post-proceso.

Redes Bayesianas:

Es un modelo probabilístico que relaciona un conjunto de variables aleatorias, muestra gráficamente redes sin ciclos representando variables aleatorias y relaciones de probabilidad que existan entre ellas que permitan obtener soluciones a problemas de decisión.

Dando a conocer un ejemplo muy simple sobre la funcionalidad de una red bayesiana. Se considera simplemente una variable aleatoria Z dependiente de otras que son F1 y F2. El grafo se representa de la siguiente forma (Lozano, 2011, p.2).

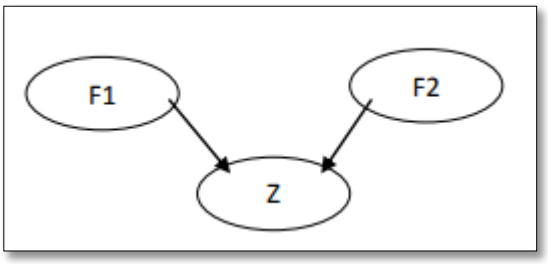


Figura 11. Grafo. Lozano (2008).

NBTree:

Es un algoritmo híbrido de Naive Bayes y Decisión Tree. Su conocimiento aprendido es representado en forma de un árbol construido recursivamente. Para atributos valorados discretos, el método Naive Bayes funciona muy bien y su rendimiento también mejora con el aumento de los datos (Malviya y Umrao, 2014, p.36).

BFTree:

Los árboles de decisión Best-First realizan la mejor división en el árbol basada en algoritmos de esfuerzo que se usa para expandir los nodos en el mejor primer orden. BFTree usa ganancia de información o índice de Gini para calcular el mejor nodo del árbol en cada paso. El mejor nodo es el que reduce en gran medida la impureza entre todos los nodos no terminales que están disponibles para la división (Srivastava y Joshi, 2014, p.729).

A continuación, se presenta un cuadro comparativo de los algoritmos de minería de datos.

Tabla 7

Cuadro comparativo de algoritmos de minería de datos

Algoritmo de predicción	Utiliza (Top Down Induction Trees)	Valor continuo	Permite ejemplos con valores desconocidos	Utiliza métodos de división	Método de Poda	Aplicación
J48	✓	✓	✓	✓	Post-Data	Aprendizaje exhaustivo
J48Graft	✓	✓	✓	✓	Post-Data	Aprendizaje exhaustivo
BFTree	✓	✓		✓	Costo Superior	Arboles de decisión
NBTree	✓	✓		✓	Costo Superior	Arboles de decisión
C45	✓	✓		✓	Post-Data	Arboles de decisión
Redes Bayesianas	✓			✓	Poli árboles	Arboles de decisión

Fuente: Vega, Rosano, López, Cendejas y Ferreira (2012).

2.2.2.5. Metodologías para implementar Minería de datos

- **Metodología KDD**

  El Descubrimiento de Conocimiento en Bases de Datos (KDD Knowdlege Discovery in Databases) constituye el primer modelo que define el descubrimiento de patrones en las bases de datos como un proceso. Este proceso tiene distintas fases que empieza desde la preparación de los datos hasta la interpretación y expansión de los resultados.

  Es una metodología propuesta por Fayyad en 1996 define a KDD como el proceso importante para identificar patrones válidos, novedosos, potencialmente útiles y entendibles en los datos.

  KDD es un proceso iterativo e interactivo. Iterativo ya que se debe seguir paso a paso y si hay alguna fase que no se completado se repite los pasos anteriores y se puede repetir muchas veces para extraer conocimiento de alta calidad. Es interactivo porque el usuario, o experto en el campo del problema, debe ayudar a la preparación de los datos y aprobación del conocimiento extraído.

Fases de la metodología KDD:

Fayyad (1996) propone 5 fases: Selección, preprocesamiento, transformación, minería de datos y evaluación e implantación.

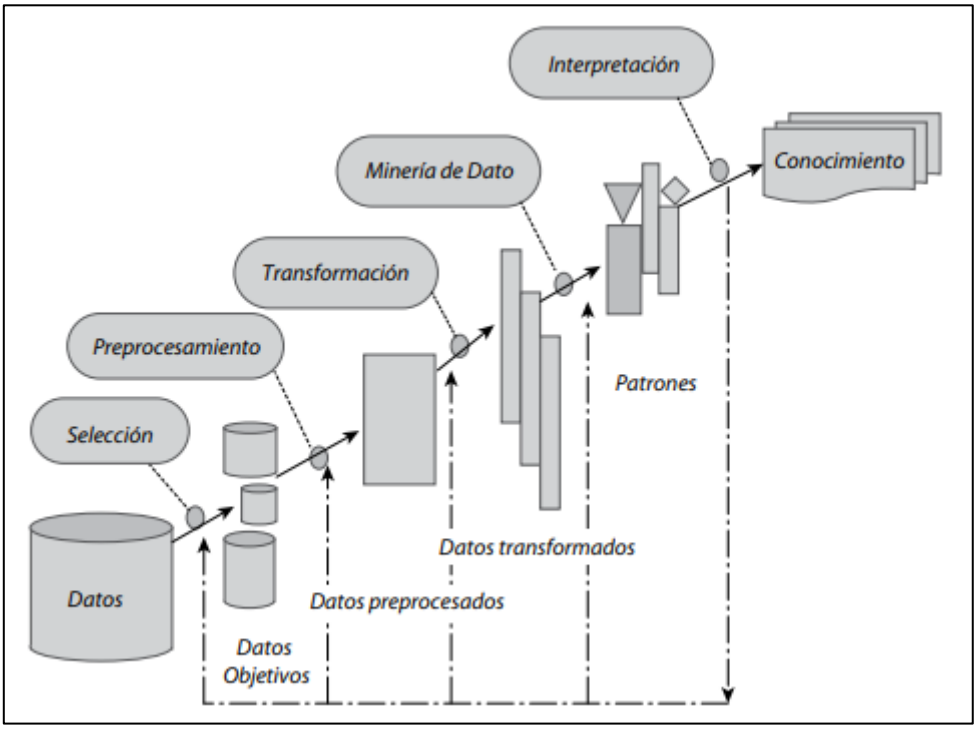


Figura 12. Fases de la metodología KDD. Timarán, Hernández, Zambrano, Hidalgo y Alvarado (2016)

- **Selección de datos:** El primer paso de la metodología KDD es la selección de los distintos orígenes que pueden tener los datos necesarios para el correcto desarrollo del modelo. Los datos deben ser seleccionados de todas las fuentes posibles.
- **Pre - procesamiento de datos:** El siguiente paso involucra limpiar los datos, siendo en este paso donde suele gastarse la mayor cantidad del tiempo del proyecto ya que los datos de múltiples fuentes suelen estar incompletos y con inconsistencias.
- **Transformación de datos:** Los modelos estadísticos utilizados para la minería de datos suelen tener requerimientos en cuanto a que tipo de datos aceptan. Esto debe ser tomado en consideración y es en esta fase donde se debe decidir para cada atributo cual será la transformación que mejor se adapta a las necesidades del modelador.

- **Minería de datos:** El siguiente paso es aplicar el modelo estadístico utilizado, que suele ser en realidad una aplicación iterativa del modelo inicialmente definido o la prueba de distintos modelos para encontrar el de mejor funcionamiento en el problema particular.
- **Interpretación y evaluación**: Este proceso considera el utilizar los resultados del modelo para crear el nuevo conocimiento, es aquí donde se analizan las secciones de implementación y resumen de la información o diseño de los entregables. El último paso de la metodología KDD es la comprensión de los resultados del proyecto.

- **Metodología CRISP-DM**

La metodología CRISP-DM está estructurada en su ciclo de vida en seis fases cada una conformada por una serie de tareas que interactúan entre sí para el desarrollo óptimo del proyecto.

Esta metodología permite tener una comprensión de los datos y prepararlos para el modelado. Es un proceso jerárquico es decir va de lo general a lo particular, algunas fases son bidireccionales, por consiguiente, se puede revisar totalmente las fases anteriores. (Contreras, Ferreira y Valle, 2017).

Fases de la metodología CRISP-DM

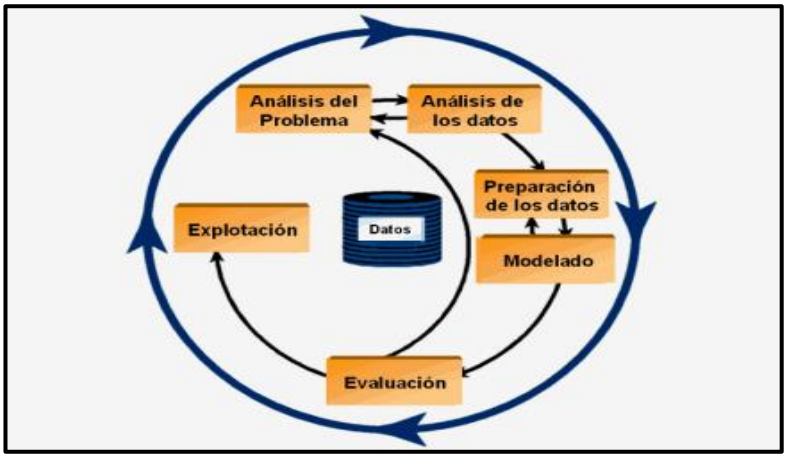


Figura 13. Fases de la metodología CRISP-DM. Vallejo y Tenelanda (2012).

1. Comprensión del negocio:
   - ✓ En esta fase se conoce los procesos relacionados con el área a tratar, es decir se determina la situación actual de la empresa. Se define los objetivos y requerimientos del proyecto. Por último, se describen los beneficios que tendrá la empresa o cliente de manera general (Contreras, Ferreira y Valle, 2017).
2. Comprensión de los datos
   - ✓ Se recolecta un conjunto de datos luego se explora los diferentes datos que existen reconociendo las características de calidad de estos, así como también sus fortalezas que sirven en el proceso de análisis.
3. Preparación de Datos
   - ✓ En esta fase se analizan los datos importantes, para ello se realiza una selección, depuración y transformación de los datos para adecuarlos a la técnica de minería de datos.
4. Modelamiento
   - ✓ Se eligen la técnica de modelado y se aplican. Se implementa las herramientas de Minería de Datos.
5. Evaluación
   - ✓ Se analizan los resultados y comportamiento de estos con la finalidad de conocer si coinciden con los objetivos del negocio.
6. Despliegue
   - ✓ Es la fase de implementación de los modelos o resultados anteriormente seleccionados, el cual el cliente usará. Desplegar los modelos resultantes en la práctica y se traza una estrategia de monitoreo del proceso.

- **Metodología SEMMA**

La metodología SEMMA (Sample, Explore, Modify, Model, Assess) fue propuesta por SAS Institute, el cual la define como el proceso de selección, exploración y modelado aplicado a cantidades

significativas de datos almacenados que permitan el descubrimiento de patrones como herramientas de apoyo para el negocio.

SEMMA se compone de cinco fases representando las etapas de un proyecto de minería de datos.

Fases de la metodología SEMMA:

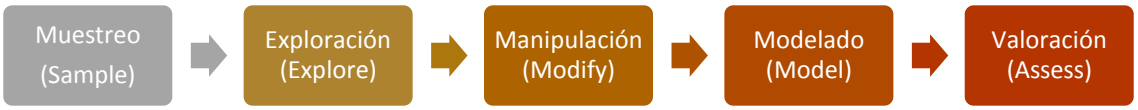


Figura 14. Fases de metodología SEMMA. Rodríguez, Álvarez, Meza, Gonzales (2003).

- Muestreo: Extracción de la población en muestra donde se aplica el análisis.
- Exploración: Exploración de los datos de la muestra con el objetivo de mejorar la eficiencia del modelo.
- Modificación: Modificación de los datos, de manera que tenga concordancia con el modelo.
- Modelado: Modelado de los datos para que el nivel de confianza sea certero.
- Evaluación: Valoración de los datos obtenidos, para contrastar con otros métodos.

- **Metodología Catalyst**

  Es una metodología propuesta por Dorian Pyle en el año 2003 y también conocida como P3TQ (Product, Place, Price, Time, Quantity). Es una metodología que formula dos modelos: modelo de negocio y el modelo de explotación de la información.

  ✓ Modelo de negocio (MII).

  Proporciona pasos para poder identificar un problema y los requerimientos reales de la organización. Contempla diferentes

ámbitos para el proyecto de minería de datos, dando a conocer acciones específicas según el escenario desde el cual se parte.

- ✓ Modelo de explotación de información (MIII).
  Proporciona una guía paso a paso para la construcción y ejecución de modelos de minería de datos a partir del modelo de negocio.

  La metodología Catalyst tanto en el modelo de negocio (MII) como en el modelo de explotación de información (MIII) están representadas en pasos, estos pasos son denominados “boxes”, lo cual estos pasos o estos saltos dependen de las situaciones que se van dando conforme se avanza el proyecto.

Fases de la metodología Catalyst

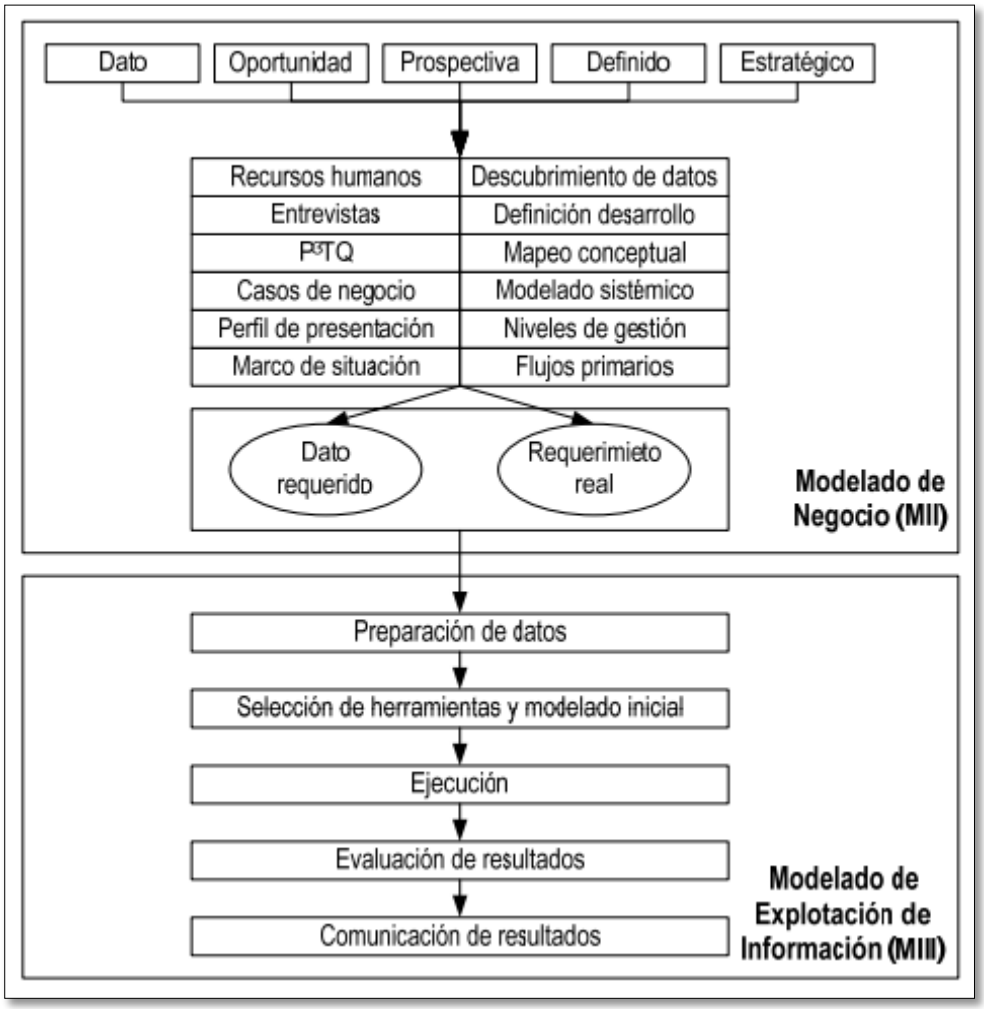


Figura 15. Interacción de los diferentes modelos (MII – MIII). Fuente: Britos (2008).

- **Comparación de las metodologías de minería de datos**

En la siguiente tabla se muestra una comparación de las diferentes metodologías para implementar minería de datos, se realizó una comparación mediante las fases de cada una de ellas.

Tabla 8

Comparación entre las metodologías KDD, CRISP-DM, SEMMA y CATALYST

Fases	KDD	CRISP-DM	SEMMA	CATALYST
Análisis y comprensión del negocio	Comprensión del dominio de aplicación	Comprensión del negocio		Modelado del negocio
Selección y preparación de los datos	Crear el conjunto de datos Limpieza y pre-procesamiento de los datos Reducción y proyección de los datos	Entendimiento de los datos Preparación de los datos	Muestreo Comprensión Modificación	Preparación de los datos
Modelado	Determinar la tarea de minería Determinar el algoritmo de minería Minería de datos	Modelado	Modelado	Selección de herramientas y modelado inicial
Evaluación	Interpretación	Evaluación	Valoración	Refinamiento del modelo
Implementación	Utilización del nuevo conocimiento	Despliegue		Comunicación

Fuente: Moine, Gordillo, Haedo (2011, p.936).

Para el desarrollo de esta investigación se pudo aplicar tanto la metodología CRISP-DM como Catalyst ya que mantienen una perspectiva más completa con respecto a los objetivos empresariales mostrando mayor completitud presentadas anteriormente en la tabla 8, mientras la metodología KDD y SEMMA están más centrados en las características técnicas del desarrollo del proceso de minería de datos proponiendo fases generales y no incorporando actividades enfocadas a la gestión del negocio evidenciando que estos modelos están orientados a aspectos técnicos. Es por ello que se optó en utilizar la metodología CRISP-DM siendo más completa y una de las metodologías más utilizadas en la actualidad en el campo de la minería de datos.

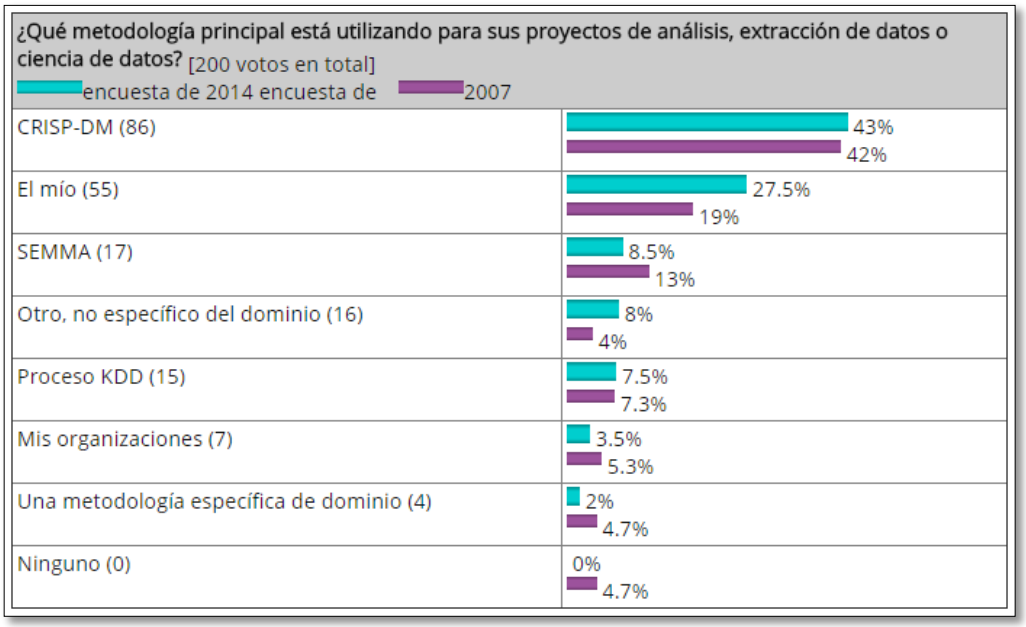


Figura 16. Metodologías más usadas. KDNuggets (2014).

.

2.3. DEFINICIÓN CONCEPTUAL DE LA TERMINOLOGÍA EMPLEADA

Riesgo: Es la posibilidad de impago de una operación. En cualquier operación existe un riesgo, para ello se debe tomar una serie de medidas para controlarlo o reducirlo.

Predicción: Utilizar algunas variables o campos en una base de datos para predecir valores desconocidos o futuros.

Algoritmo: Según Baños y Hernández (2012) define que un algoritmo “es un conjunto de pasos lógicos y estructurados que nos permite dar solución a un problema”. Es decir, es un conjunto de heurísticas y cálculos que permiten crear un modelo a partir de datos.

Entropía: Numero de bits necesarios para codificar un suceso. Cuantos más bits, más información menos probable es un suceso.

Ganancia: La ganancia es la información del conjunto de datos es decir que cuanto mayor sea, la información que aporta será menor, es decir es probable que sea un buen candidato como atributo importante del conjunto.

2.4. ESTADO DEL ARTE

En términos generales la evolución de la minería de datos inicia cuando se requiere almacenar la información de las empresas en las computadoras. Es ahí donde la minería de datos va más allá de solo almacenar sino también analizar la información para luego tomar decisiones y mostrar resultados. La minería de datos es un conjunto de técnicas de análisis que permiten extraer patrones, tendencias para describir y comprender mejor los datos.

Esta herramienta ha sido desarrollada por varias disciplinas como la visualización, inteligencia artificial, estadística y las tecnologías de base de datos. La evolución de las herramientas del DM (Data Mining) en el transcurso del tiempo se divide en cuatro etapas principales: Colección de datos (1960), Acceso de datos (1980), Almacén de datos y apoyo a las decisiones (principios de la década de 1990) y Minería de datos inteligente (finales de la década de 1990).

Tiempo atrás las empresas no se interesaban por tener una estrategia comercial para incrementar sus ventas, disminuir costos, identificación de clientes que pasarían a la competencia, predecir qué cantidad de productos comprará un cliente, etc. Todas estas estrategias no se podrían llevar a

cabo sin la minería de datos. La minería de datos automatiza estos procesos reduciendo los tiempos hasta llegar a una decisión.

2.4.1. Árboles de decisión

Los árboles de decisión aparecen a mediados del año 1960. Son una representación de sucesos o evento de una manera gráfica para una mejor y fácil interpretación. Según Bouza y Santiago (2012) mencionan que: “Los árboles de decisión proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos los resultados por cualquier usuario”.

El uso de la técnica de árboles de decisión tiene distintas ventajas entre ellas se encuentran la fácil interpretación para una mejor toma de decisión, proporciona diferentes soluciones a un problema y permite analizar las consecuencias antes de tomar una decisión.

El árbol de decisión está compuesto por nodos; nodo de decisión, nodo de probabilidad, nodo terminal u hoja y rama.

- ✓ **Nodo de decisión**: Este nodo indica que una decisión necesita tomarse en ese punto del proceso
- ✓ **Nodo de probabilidad:** Este nodo indica que sucede un evento aleatorio, es decir tiene varias divisiones.
- ✓ **Nodo terminal u hoja:** Este nodo indica que ya no es posible dividir.
- ✓ **Rama:** Indica los caminos para tomar una decisión o ya sea un evento aleatorio.

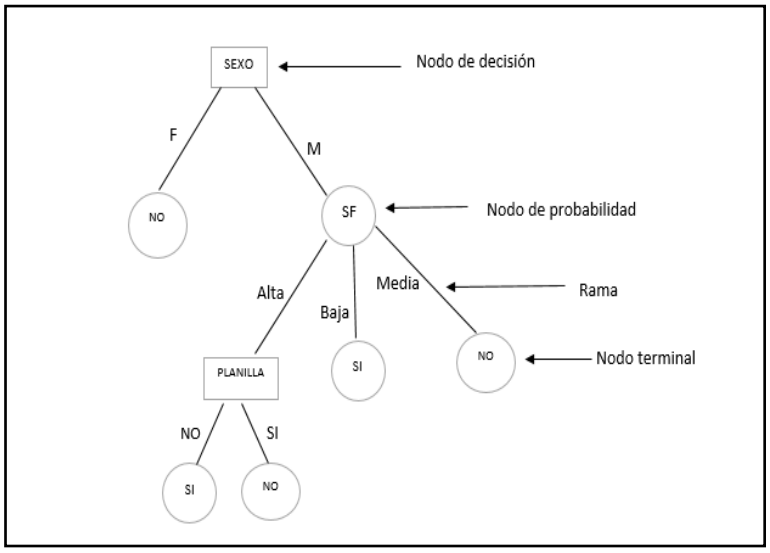


Figura 17. Ejemplo de predicción de riesgo de morosidad basado en la terminología.

2.4.2. Algoritmos de árboles de decisión

Los algoritmos más utilizados fueron creados hace 30 años, lo que hace que hoy en día existan herramientas para minería de datos que generan resultados con alta confiabilidad.

Para la técnica de árboles de decisión se utilizan diferentes algoritmos, entre ellos se encuentran: ID3, C4.5, J48, Redes bayesianas, NBTree, decisión tree, CART, etc.

2.4.3. Usabilidad del algoritmo ID3

Existen diferentes estudios que aplican árboles de decisión a problemáticas similares al estudio, sin embargo, se ha tomado en cuenta el algoritmo ID3 basado en las referencias en cuanto a la precisión de dicho algoritmo. A continuación, se describen algunos trabajos previos:

Solarte y Soto (2011) en su artículo “Arboles de decisión en el diagnóstico de enfermedades cardiovasculares”, plantean el uso de árboles de decisión para determinar si es o no posible administrar fármacos a pacientes con enfermedades cardiovasculares de un hospital. Analizaron variables como la presión arterial, azúcar en la sangre, índice de colesterol, alergias a antibióticos, otras alergias y administrar fármacos. Utilizaron el algoritmo ID3 de árboles de decisión. Una vez analizadas las variables y construcción de la base de datos, pasaron al modelo

matemático del algoritmo ID3. Como resultado de la formula obtuvieron que el atributo presión arterial tuvo una mayor ganancia, el cual paso como nodo o raíz del árbol a construir, y así sucesivamente realizaron el mismo procedimiento para los demás atributos hasta obtener el árbol. Como resultado de esta investigación se obtuvo que sí es posible administrar fármacos a pacientes con dichas enfermedades mediante la técnica de árboles de decisión.

Por otra parte, los investigadores colombianos Tello, Eslava y Tobías (2012) realizaron un estudio donde se evaluó el nivel de riesgo en el otorgamiento de créditos de una cartera comercial de una entidad bancaria. Aplicaron la técnica de minería de datos y realizaron una comparación de los algoritmos ID3 y J48. Se obtuvo diferentes atributos como edad de mora, saldo a cápita, modalidad, comportamiento de pago, etc. Estos datos fueron analizados con la herramienta de minería de datos WEKA, el cual arrojo como resultado que el algoritmo ID3 tuvo mejor instancia correcta que el algoritmo J48, la cual quiere decir que tuvo una mejor precisión para el problema.

Otro estudio, desarrollado en la India por Bhatt, Mehta y D'mello (2015), da cuenta del uso del algoritmo de árbol de decisión ID3 para la predicción de colocación, tuvo como objetivo determinar o identificar las cualidades más relevantes que puede tener un estudiante en la universidad de Mumbai en la India, identificando el perfil más adecuado para incorporarse a un puesto de trabajo. Para ello se diseñó un modelo predictivo que ayude a predecir la colocación del estudiante. Entre las variables más importantes que se tuvieron en cuenta en esta investigación fue el rendimiento académico, habilidades técnicas y comunicación y por último habilidades de programación. Esta investigación tuvo como prioridad ayudar a los planificadores académicos para diseñar estrategias en la universidad de Munbai orientado hacia los estudiantes y a su vez mejorar su rendimiento académico para colocarlos en un puesto de trabajo en el menor tiempo posible. Por último, se concluyó que la utilización de los

algoritmos de clasificación puede emplearse con éxito para la predicción de colocación para un puesto de trabajo.

Finalmente, el estudio desarrollado en Nigeria por Adeyemo y Adeyeye (2015) titulado “Estudio comparativo de algoritmos de perceptron de árbol de decisión ID3/ C4.5 y multicapa para la predicción de la fiebre tifoidea”, tuvo como objetivo realizar una comparación del desempeño de los algoritmos predictivos. Los algoritmos que utilizaron para el problema fueron extraídos del hospital de Nigeria utilizando algoritmos redes neuronales, algoritmo ID3 y C4.5 en la herramienta WEKA. En esta investigación se pretendió reducir la tasa de mortalidad por causa de la fiebre tifoidea. Entre los atributos están la edad, sexo, dolor abdominal, dolor de cabeza, etc. Como resultado se obtuvo que las redes neuronales tienen una mayor precisión que los otros algoritmos de clasificación.

CAPÍTULO III
DESARROLLO DEL SISTEMA

3.1. ESTUDIO DE FACTIBILIDAD

3.1.1. Factibilidad técnica

Para el desarrollo de esta investigación es necesario conocer la información de las características físicas y técnicas de cada elemento que se utilizaron en este estudio con la finalidad de poner en funcionamiento el desarrollo del proyecto.

A continuación, se describen los aspectos técnicos que se deben contemplar para la implementación del proyecto.

Tabla 9

Software disponible

RECURSOS	DESCRIPCIÓN	CARACTERÍSTICAS
HARDWARE	➢ Disco duro ➢ Memoria RAM ➢ Procesador ➢ Mouse ➢ Teclado	➢ 1TB ➢ 8GB (instalados) / Max 32GB Tipo DDR3 ➢ Intel(R) Core(TM) i5-6200 CPU @2.30GHz (4 CPUs) ➢ DATAONE ➢ DATAONE
SOFTWARE	➢ Sistema Operativo ➢ Microsoft Office ➢ Netbeans ➢ Servidor Web ➢ Xampp / Wampserver	➢ Windows 10, 8 x64 ➢ 2016, 2013 ➢ v8.0, v8.2 ➢ Apache tomcat v8.5.31 / GlassFish Server v4.1.1 ➢ v3.2.2 / v2.4
OTROS	➢ Conexión a Internet	➢ 8MBps

3.1.2. Factibilidad operativa

El proyecto de investigación es viable operativamente, porque el personal encargado del área del departamento de cobranza que dará uso del sistema Web tendrá una previa capacitación con el fin de no tener inconvenientes al momento de utilizarse, además se cuenta con el apoyo

del jefe del departamento de cobranza teniendo la experiencia necesaria en el tema de minería de datos razón por la cual da mayor facilidad y mejor orientación con lo que corresponde al proyecto.

3.1.3. Factibilidad económica

El desarrollo del proyecto es viable económicamente, ya que la empresa dispone de los recursos económicos necesarios para la implementación del proyecto. Esta inversión ayudará a tomar mejores decisiones en el departamento de cobranza aplicando minería de datos con el fin de tomar mejores estrategias enfocados a los clientes.

Tabla 10

Costos de desarrollo de la solución

RECURSOS GENERALES	DESCRIPCIÓN	TIPO DE UNIDADES	COSTO (S/.)	TOTAL (S/.)
		Recursos Humanos		
Espino Quiñones Leonardo	Persona	1 Und.	S/3,000.00	S/3,000.00
Garcia Torres Maria Emily	Persona	1 Und.	S/3,000.00	S/3,000.00
		Recursos Hardware		
Impresora multifuncional	HP Deskjet 3636 Aio	1 Und.	S/600.00	S/600.00
Laptop	Lenovo Intel Core i5 2.4GHz(4CPUs)	1 Und.	S/2,200.00	S/2,200.00
USB	HP 16Gb	2 Und.	S/35.00	S/70.00
		Recursos Software		
Windows 10	S.O de 64bits	1 Und.	S/900.00	S/900.00
Microsoft office	2016	2 Und.	S/550.00	S/1,100.00
Bizagi Modeler	v2.9.0.4	2 Und	Free	-
Netbeans	v8.0, v8.2	2 Und	Free	-
		Otros Gastos		
Impresiones	Tesis	Und.	S/400.00	S/400.00
Materiales de oficina	-	Und.	S/60.00	S/70.00
Total Presupuesto				**S/11,340.00**

3.2. APLICACIÓN DE LA METODOLOGÍA CRISP-DM

3.2.1. Comprensión del Negocio

A continuación, se da a conocer cada una de las tareas en la fase de comprensión del negocio correspondiente a la metodología CRISP-DM, cuyo fin es determinar los objetivos del proyecto desde una perspectiva empresarial o institucional y generando un plan preliminar diseñado para alcanzar dichos objetivos.

3.2.1.1. Determinar el objetivo del negocio

- Departamento de Telemarketing

La empresa de seguros Oncosalud actualmente cuenta con el departamento de telemarketing encargado de realizar las ventas de seguros Oncológicos desde un Call Center (empresa terciaria) a todas las personas que requieran de este servicio, estas ventas realizadas hacia el cliente son consideradas como altas siempre y cuando el cliente haya pagado la primera cuota de su seguro Oncológico de lo contrario no será considerado como alta o venta aprobada.

Seguidamente, todas las llamadas aprobadas o consideradas como altas pasan a un proceso de inscripción donde se registran a todos los clientes que adquieren su seguro oncológico y posteriormente llevar a cabo el tratamiento que le corresponde a cada uno de ellos.

- Departamento de cobranza

Por otro lado, el departamento de cobranza de igual forma debe realizar las llamadas desde un Call Center a todas las personas que se les ha vendido un seguro Oncológico para sostener un trato directo con sus clientes y mantenerlos informados constantemente respecto a su cuota por vencer o cuotas vencidas.

En la actualidad el departamento de cobranza no cuenta con una herramienta que le ayude a identificar o a detectar a los clientes más propensos a caer en morosidad, lo cual hace que tomen decisiones con

menor grado de certeza y no se utilice una gestión preventiva respecto sus clientes.

A continuación, se muestra una representación gráfica respecto a la situación actual desde que se realiza la venta de un seguro Oncológico hasta el cobro del seguro del cliente.

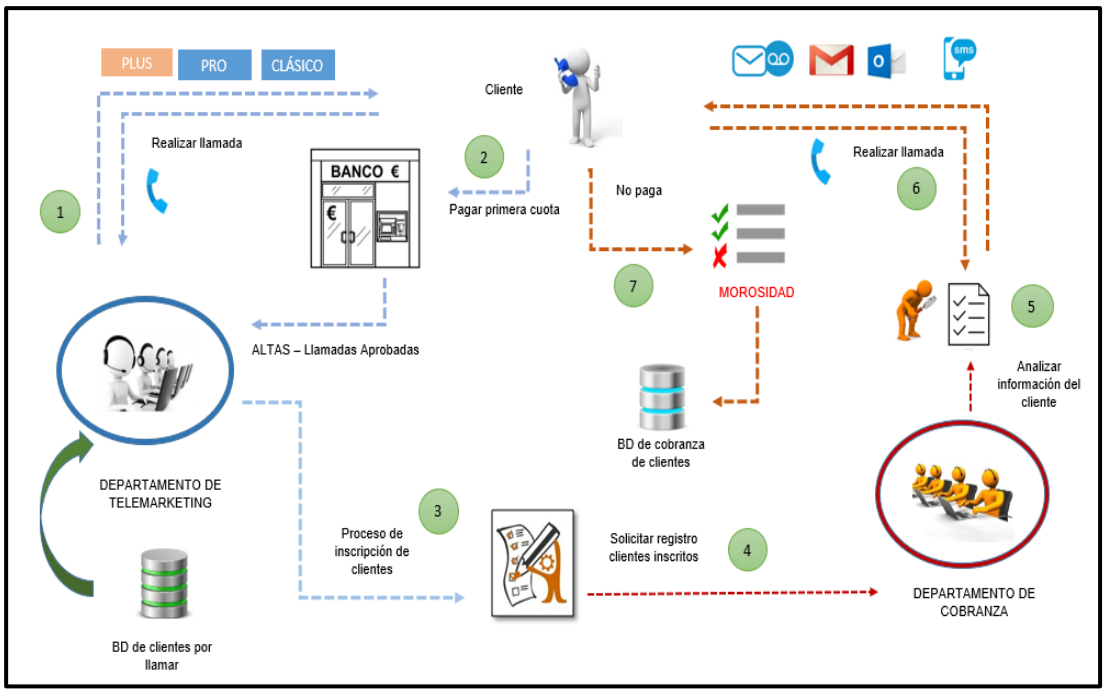


Figura 18. Situación actual del negocio

Objetivos del negocio

En cuanto a los objetivos que tiene el departamento de cobranza son los siguientes:

- Gestionar la recuperación oportuna de las cuentas por cobrar a los clientes.
- Plantear estrategias de cobro y proponer soluciones con los métodos más adecuados para evitar que las cuentas por cobrar no corran el riesgo de caer en morosidad alta.
- Evitar que la morosidad sea mayor que la presupuestada.

Criterios de éxito del negocio

Desde el enfoque de negocio, se establece como criterio de éxito los siguientes indicadores:

Tabla 11

Indicadores para medir el criterio de éxito del negocio

	Indicadores	Medidas
Indicador	Índice de morosidad	$\frac{N° \ clientes \ atrasados \ en \ sus \ cuotas}{N° \ clientes \ que \ deben \ pagar \ sus \ cuotas}$
	Porcentaje de contención de clientes	$\frac{N° \ clientes \ contenidos}{N° \ clientes \ que \ iniciaron \ el \ mes}$
	Porcentaje de recupero	$\frac{N° \ clientes \ recuperados \ en \ sus \ pagos}{N° \ de \ clientes \ que \ iniciaron \ el \ mes \ con \ mora}$

3.2.1.2. Evaluación de la situación

Para la realización del proyecto de minería de datos se cuenta con una base en Microsoft Excel que brinda la información detallada de los clientes que fue facilitado por el jefe de recupero de clientes del departamento de cobranza, por lo que da mayor aporte a este estudio y poner en marcha el desarrollo de esta investigación.

Inventario de recursos

En cuanto a los recursos disponibles para el desarrollo del proyecto, contamos con las siguientes herramientas:

- ✓ **Weka** (v3.6.15). Es una herramienta para el aprendizaje automático y minería de datos diseñado en java en la universidad de waikato, además es una herramienta de distribución de licencia GNU-GLP o software libre.
- ✓ **MySql** (v4.7.9). Es un sistema de gestión de base de datos o SGBD que permite ser usado por varias personas al mismo tiempo. Fue

desarrollado en C y C++ y se destaca por su gran adaptación a diferentes entornos de desarrollo.

La fuente de datos para la extracción de la información de los clientes de la empresa se encuentra ubicada en una base histórica en Microsoft Excel.

Requerimientos

- Se requiere de una herramienta que facilite la predicción de los clientes más propensos a caer en morosidad a partir de datos de datos históricos de los clientes de la empresa.
- Se requiere de una herramienta que muestre información rápida y visual sobre el riesgo de morosidad del cliente ya sea alto, medio o bajo.

Costes y beneficios

Los datos que se utilizaron para el desarrollo del proyecto no generaron ningún coste adicional al proyecto ya que los datos fueron proporcionados por la misma empresa.

En cuanto a los beneficios del proyecto, se puede afirmar que el proyecto sí genera un beneficio económico a la empresa Oncosalud, puesto que el objetivo del proyecto es predecir el riesgo de morosidad de los nuevos clientes, lo cual permitirá tomar acciones anticipadas y así poder reducir el índice de morosidad de la empresa.

3.2.1.3. Determinar los objetivos del proyecto

A partir de la situación actual y los requerimientos planteados por el departamento de cobranza se plantea los siguientes objetivos del proyecto.

Objetivo general

Predecir el riesgo de morosidad de los clientes en base al modelo predictivo que ayudará a identificar que clientes tendrán un mayor riesgo a futuro, con la finalidad de tomar decisiones con mayor determinación en el departamento de cobranza.

Objetivos Específicos

- Llevar un control de la información de los posibles clientes morosos para aplicar estrategias en base a los resultados obtenidos.
- Identificar información relevante mediante gráficos estadísticos.

3.2.1.4. Realizar un plan de proyecto

El proyecto se dividió en las siguientes fases para poder estimar el tiempo de ejecución:

Tabla 12

Plan del proyecto

Fase	Tiempo estimado
Análisis de los objetivos del negocio y los criterios de éxito.	2 semanas
Análisis de la estructura de los datos y la información de la base de datos.	3 semanas
Preparación de los datos (Selección de datos, limpieza y conversión) para la minería de datos.	4 semanas
Elección de la técnica de modelado y ejecución sobre los datos.	3 semanas
Análisis de los resultados obtenidos.	2 semanas
Producción de los informes con los resultados obtenidos y resultados finales.	1 semana

3.2.2. Comprensión de los datos

Esta fase comprende la recolección de los datos iniciales para el proyecto de investigación con la finalidad de mantener un contacto con el problema y familiarizarse con los datos recolectados, teniendo en cuenta siempre los objetivos del negocio.

3.2.2.1. Recolectar los datos iniciales

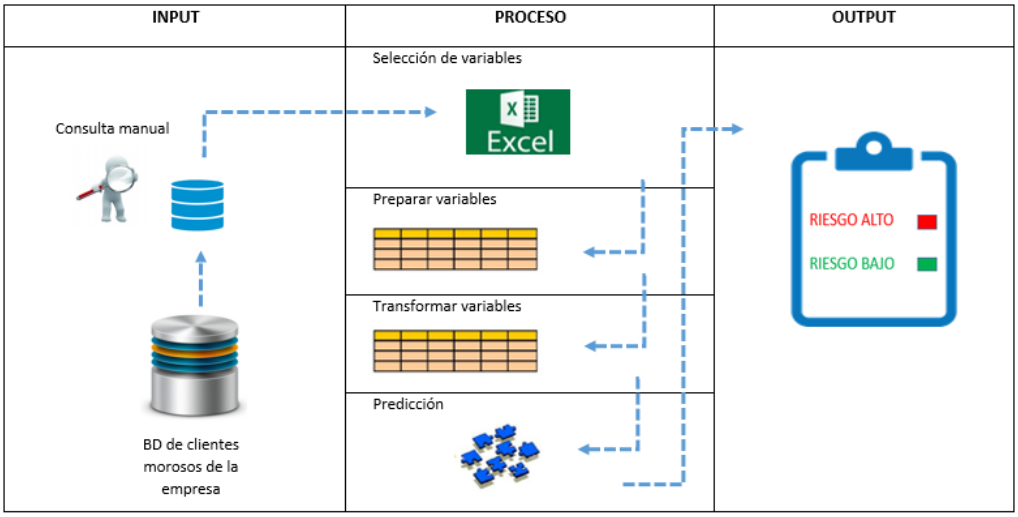
Los datos utilizados en el desarrollo del proyecto son datos históricos de los clientes facilitado por la misma empresa Oncosalud, de manera que se tomó la información que presento mayor relevancia para esta

investigación que incluye el sexo, edad, fecha de inscripción, estado civil, ocupación, tipo de tarjeta, programa, deuda en Oncosalud, categoría, calificación crediticia, entre otros. Es importante recalcar que la información utilizada para el estudio son datos reales que se encuentran almacenados en la base de datos de la misma empresa.

A continuación se muestra una representación gráfica del proceso de recolección de los datos desde la entrada de la información (INPUT) para la selección de las variables hasta llegar al conocimiento siendo la salida de la información (OUTPUT).

Tabla 13

Input y output de la recolección de los datos



En cuanto a la transformación de los datos, se ha tenido que convertir algunos atributos numéricos ya que el algoritmo que se va a utilizar solo puede ser utilizado con datos categóricos.

Entre ellos se hizo el cambio correspondiente a las variables que son: edad, tiempo de trabajo, ingreso mensual y número de hijos para la integración con el algoritmo de clasificación.

3.2.2.2. Descripción de los datos

Los datos utilizados para el estudio de investigación se encuentran almacenados en una base de datos de los cuales se ha realizado una

descripción con su respectivo tipo de datos de las variables que se muestran en la siguiente tabla:

Tabla 14

Diccionario de datos

Campo	Tipo	Descripción
Cod_afi	Int	Es el atributo que representa el número de código del cliente afiliado.
Cod_gf	Int	Es el atributo que representa el código de grupo familiar es decir este código comprende a un grupo de personas.
Sexo	Categórico	Es el atributo que representa el género por cada cliente.
Estado_civil	Categórico	Es el atributo que representa la condición en la que se encuentra el cliente.
Edad	Int	Es el atributo que identifica la edad de cada cliente registrado.
Fecha_proc	Date	Este atributo representa la fecha de inicio en la que inicia el contrato del cliente.
Tipo_doc_afi	Texto	Es el atributo que representa el tipo de documento que maneja el cliente.
Nrodociden_afi	Int	Es el atributo que representa el número del tipo documento que maneja el cliente.
Dias_morosidad	Int	Es el atributo que representa los días de atraso a partir de la fecha vencida de un cliente.
Calif_crediticia	Categórico	Este atributo representa la situación financiera en la que se encuentra registrado actualmente cada cliente ya sea ante cualquier entidad.
Ocupación	Categórico	Este atributo representa la actividad económica que realiza cada cliente ya sea por cuenta propia o de manera subordinada
Planilla	Categórico	Es el atributo que demuestra si el cliente se encuentra registrado o no en planilla por el tiempo laborando.
Seg_gestión	Categórico	

		Este atributo representa el seguimiento que se tomará hacia el cliente dependiendo de los días de morosidad que presenta.
Tiempo_trabajo	Int	El atributo indica el tiempo que se encuentra laborando el cliente.
Ingreso_men	Int	Este atributo representa el dinero promedio que recibe regularmente el cliente en la empresa que se encuentra laborando.
Frecuencia	Categórico	Este atributo representa la modalidad de pago que utilizará el cliente para pagar su cuota ya sea mensualmente o anualmente.
Deuda_Onco	Categórico	Es un atributo que demuestra si el cliente ha tenido alguna deuda pagada anteriormente con la empresa de seguros Oncosalud.
Categoría	Categórico	Es un atributo que representa el tipo de parentesco que tiene el cliente ante el domicilio que reside.
Fuma	Categórico	Este atributo especifica si el cliente fuma o no fuma para definir el coste que el cliente deberá pagar por su seguro.
Num_hijos	Int	El atributo representa el número de hijos que cuenta cada cliente.
Celular1	Int	Este atributo representa el número de teléfono que maneja el cliente.
Fijo1	Int	Este atributo representa el número telefónico de casa que maneja el cliente.
Programa	Categórico	El atributo representa el tipo de programa que adquiere cada cliente para su tratamiento Oncológico.
Tipo_tarjeta	Categórico	Es el atributo que representa el tipo de tarjeta que cuenta cada cliente.
Otra_tarjeta	Int	Es el atributo que representa cuantas tarjetas adicionalmente maneja.
Riesgo	Categórico	Es el atributo que se quiere predecir (LABEL).

3.2.2.3. Exploración de los datos

Una vez que se ha definido los datos se procede al siguiente paso realizar los análisis estadísticos básicos de los datos, en el cual se crean gráficos de distribución de los datos.

En el gráfico 19 se muestra el porcentaje de clientes de la empresa con el riesgo de morosidad que han obtenido en los meses de junio, julio y agosto del 2018.

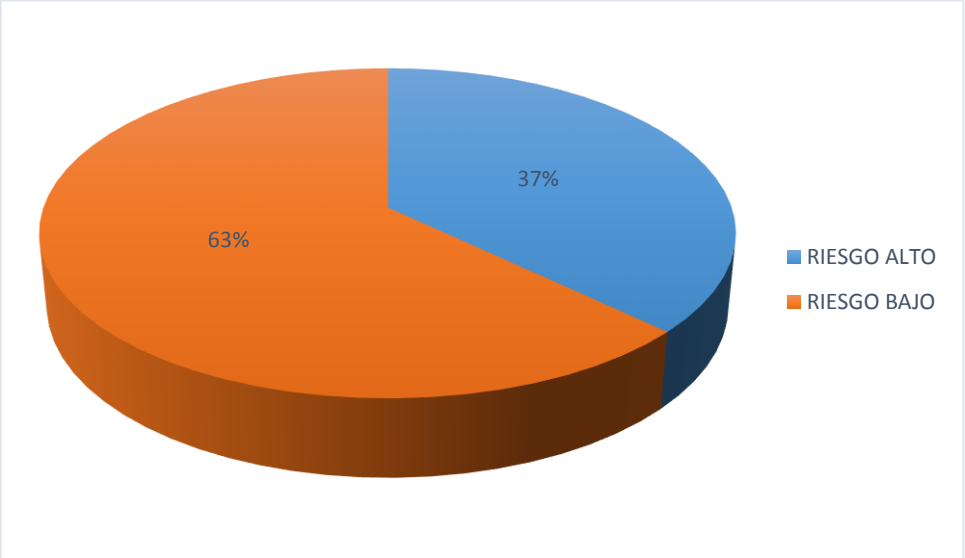


Figura 19. Porcentaje de clientes por riesgo de morosidad.

En el gráfico 20 se muestra la cantidad de los clientes de la empresa por su categoría y su riesgo de morosidad analizado por la empresa.

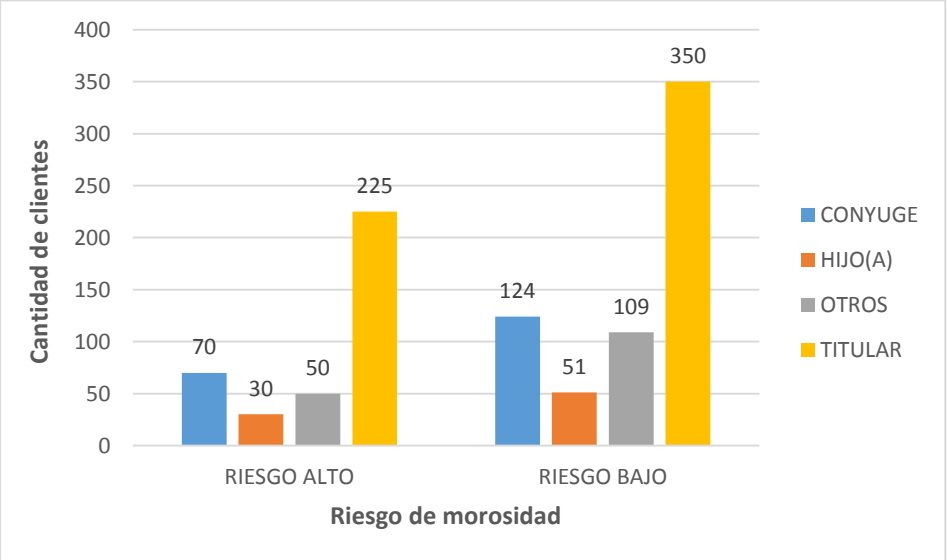


Figura 20. Cantidad de clientes por categoría y riesgo de morosidad.

En el gráfico 21 muestra la cantidad de clientes según su sexo con su respectivo riesgo de morosidad.

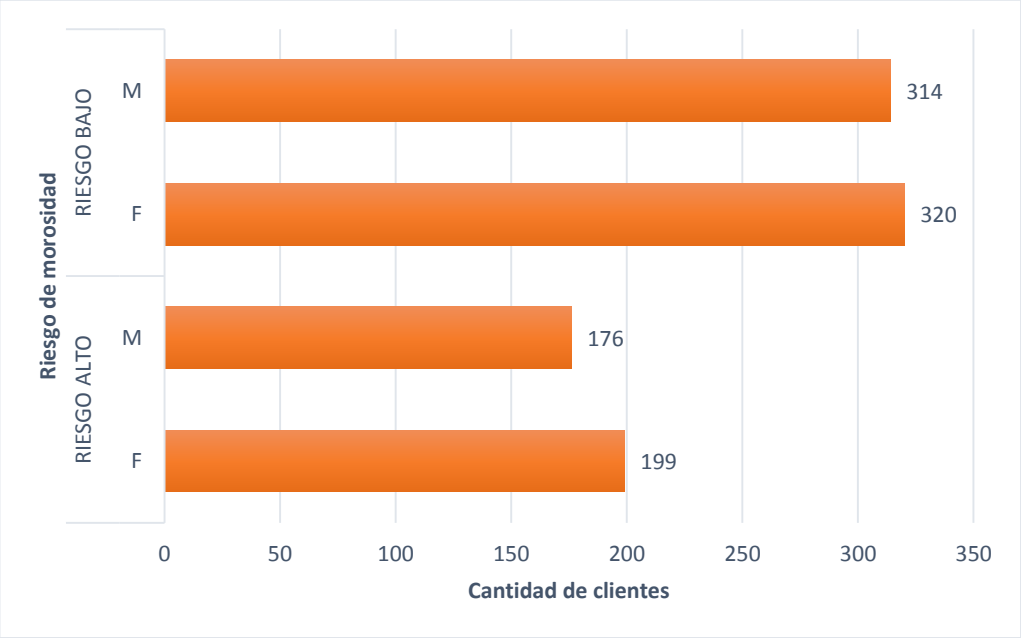


Figura 21. Cantidad de cliente según su sexo por riesgo de morosidad.

En el gráfico 22 muestra la cantidad de clientes con su respectivo programa oncológico de la empresa.

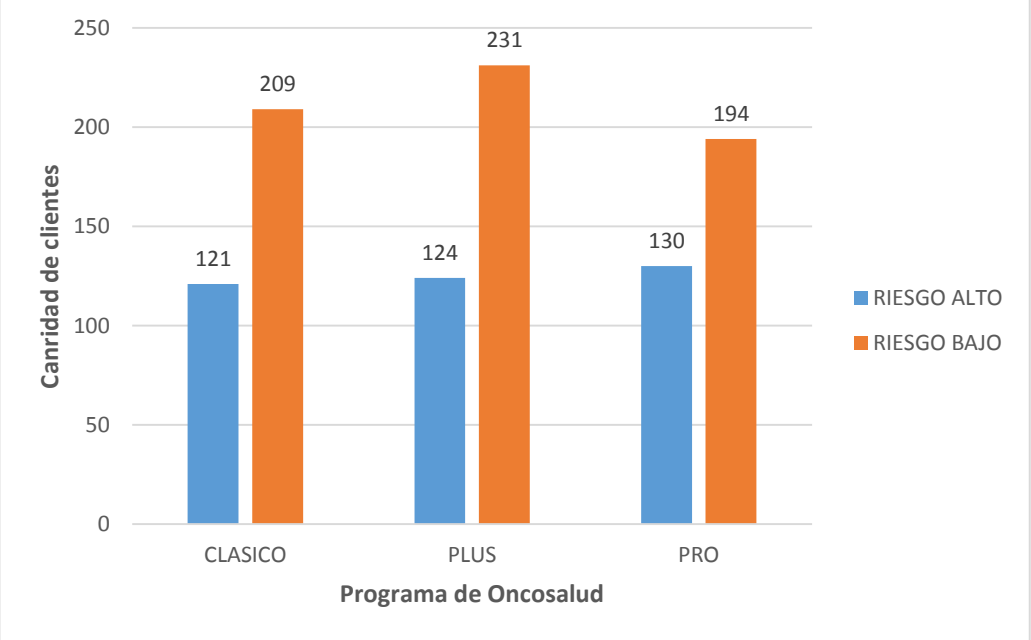


Figura 22. Cantidad de clientes por programa.

En la figura 23 muestra la cantidad de clientes mediante su calificación crediticia que obtuvo un determinado riesgo de morosidad.

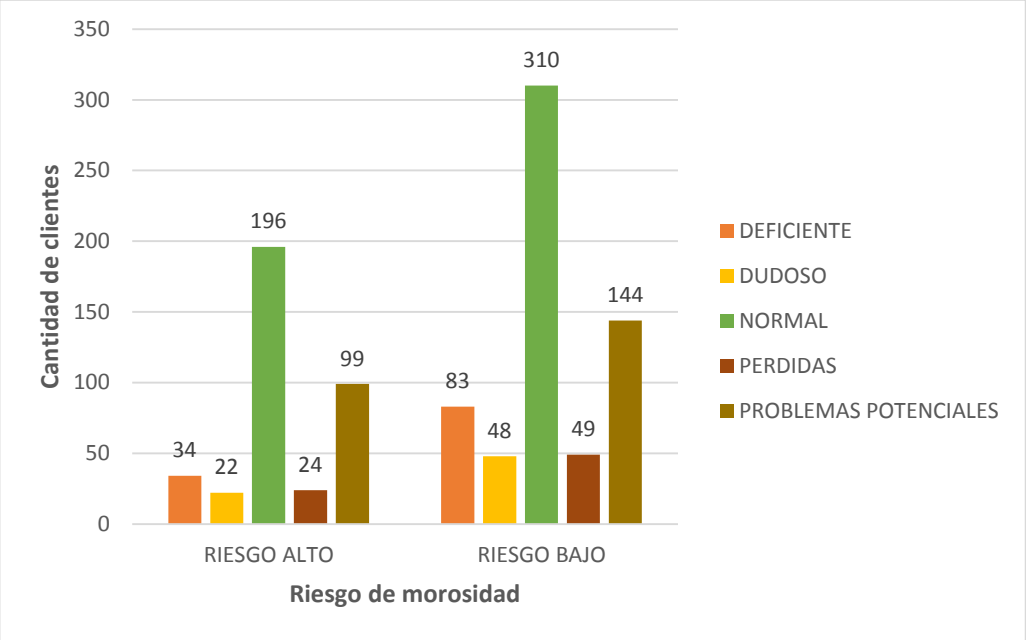


Figura 23. Cantidad de clientes con su calificación crediticia mediante el riesgo.

3.2.2.4. Verificar la calidad de los datos

El paso final en esta fase de la metodología CRISP-DM se afirma que los datos son completos para poder realizar el objetivo del proyecto, los datos que se han extraído no contienen errores, es decir no hay valores erróneos. Tampoco existen valores vacíos en ninguna variable por lo que se ha tomado todos los registros de los clientes extraídos de la empresa. Esto quiere decir que faltaría separar los datos que se requiera para el modelo lo cual se desarrollara en la siguiente fase de la metodología.

3.2.3. Preparación de los datos

En esta fase se procede a preparar los datos para ser adaptados a la minería de datos que serán utilizados posteriormente. Esto incluye las tareas generales de selección de datos, limpieza de datos, construcción de datos, integración de datos y formateo de datos que se detallan a continuación:

3.2.3.1. Seleccionar los datos

En esta etapa se utilizarán todos los registros de cada tabla que compone la base de datos, ya que ha sido creada para esta investigación. La cantidad de registros que han sido seleccionados para esta investigación fueron todos los datos proporcionados por la misma empresa.

Los atributos seleccionados para el análisis son los siguientes:

- Sexo
- Edad
- Estado civil
- Tiempo de trabajo
- Ingreso mensual
- Programa
- Deuda en la empresa por cuotas
- Categoría
- Calificación crediticia
- Número de hijos
- Riesgo de morosidad

3.2.3.2. Limpiar los datos

Los datos fueron extraídos de la base de datos de la misma empresa por lo cual no se han encontrado muchos problemas en los datos. En este proceso de limpieza de datos no ha habido valores que están fuera del rango de los atributos, ni valores nulos ni datos incoherentes o datos que deberían estar en otros atributos por lo que ha sido más sencillo realizar este proceso. En el caso de que existieran valores vacíos o nulos, al momento de realizar la minería de datos es no tomar en cuenta esos registros, ya que no ayudaría para el modelo.

3.2.3.3. Construir los datos

En este proceso se realiza la construcción de la base de datos con las variables que se necesitan para la predicción y los datos de los clientes.

En la siguiente tabla se muestran los posibles valores que puede tomar cada variable que se ha construido para desarrollar el modelo predictivo. Ver tabla N° 15.

Tabla 15

Variables objetivo del modelo

VARIABLE	DEFINICIÓN
Sexo	Sexo del cliente: a) Femenino b) Masculino
Edad	Años de edad del cliente. a) 18 a 25 años b) 26 a 35 años c) 36 a más años
Estado civil	Estado civil del cliente: a) Soltero b) Casado c) Viudo d) Divorciado
Calificación crediticia	Calificación crediticia del cliente: a) Normal b) Problemas Potenciales c) Deficiente d) Dudoso e) Pérdidas
Tiempo de trabajo	Años que está trabajando el cliente: a) Menor a 2 años (<2) b) Entre 2 y 5 años (2<=x<=5) c) Mayor a 5 años (>5)
Ingreso mensual	Sueldo promedio del cliente: a) Menor a 1000 soles (<1000) b) Entre 1000 y 3000 soles (1000<=X<=3000) c) Mayor a 3000 soles (>3000)
Deuda en Oncosalud	Deuda antigua pagada del cliente en la empresa: a) No tiene b) 1 a 2 cuotas (1<=x<=2) c) 3 a 4 cuotas (3<=x<=4) d) 5 a más cuotas (>=5)

Categoría	Categoría del cliente: a) Titular b) Conyugue c) Hijo(a) d) Otros
Número de hijos	Cantidad de hijos del cliente: a) No tiene b) De 1 a 2 (1<=x<=2) c) Entre 3 a más (>=3)
Programa	Programa del cliente: a) Clásico b) Plus c) Pro
Riesgo	Riesgo de morosidad: a) Bajo b) Alto

3.2.3.4. Integrar los datos

Para realizar este proceso no ha sido necesario agregar nuevos atributos ni nuevos campos para realizar la evaluación de la minería de datos, ya que como se ha mencionado anteriormente los campos han sido extraídos y solo se han seleccionado los atributos importantes de la base de datos de los clientes de la empresa.

3.2.3.5. Formateo de los datos

Se ha formateado algunos de los atributos con el fin que el algoritmo a utilizar pueda procesar estos datos, ya que solo recibe datos nominales. En la base de datos inicial, los atributos edad, tiempo de trabajo, sueldo promedio y número de hijos eran numéricos por lo cual se transformó a categórico con los siguientes valores que se muestran a continuación.

Para el atributo edad:

- A = 18 a 29 años
- B = 30 a 45 años
- C = 46 a más años

Para el atributo tiempo de trabajo:

- A = Menor a 2 años
- B = Entre 2 a 5 años
- C = Mayor a 5 años

Para el atributo ingreso mensual:

- A = Menor a 1000 soles
- B = Entre 1000 y 3000 soles
- C = Mayor a 3000 soles

Para el atributo deuda Oncosalud:

- A = No tiene
- B = 1 a 2 cuotas
- C = 2 a 3 cuotas
- D = 5 a más cuotas

Para el atributo número de hijos:

- A = No tiene
- B = De 1 a 2 hijos
- C = Entre 3 a más hijos

A continuación se muestra los datos que se van a utilizar para el modelo, los cuales están con las variables que se necesitan y los valores correspondientes.

SEXO	EDAD	ESTADO CIVIL	TIEMPO_TRAB	INGRESO_MENS	PROGRAMA	DEUDA_ONCO	CATEGORIA	CALIFICACIÓN	NUM_HIJOS	RIESGO
M	A	SOLTERO	C	B	PLUS	C	CONYUGE	PROBLEMAS POTENCIALES	C	RIESGO ALTO
F	A	SOLTERO	A	B	PLUS	A	TITULAR	PROBLEMAS POTENCIALES	A	RIESGO BAJO
M	A	SOLTERO	A	B	CLASICO	A	TITULAR	NORMAL	A	RIESGO BAJO
F	C	SOLTERO	A	C	CLASICO	A	OTROS	PROBLEMAS POTENCIALES	B	RIESGO BAJO
M	A	DIVORCIADO	A	C	CLASICO	A	TITULAR	NORMAL	B	RIESGO BAJO
M	B	DIVORCIADO	A	B	PRO	B	TITULAR	NORMAL	B	RIESGO BAJO
F	B	VIUDO	B	C	PLUS	A	TITULAR	NORMAL	B	RIESGO BAJO
M	C	SOLTERO	B	C	PLUS	A	CONYUGE	PROBLEMAS POTENCIALES	A	RIESGO BAJO
F	B	CASADO	B	C	PRO	A	CONYUGE	NORMAL	A	RIESGO BAJO
M	C	DIVORCIADO	C	B	PRO	A	TITULAR	NORMAL	B	RIESGO BAJO
M	C	VIUDO	A	B	PRO	C	TITULAR	NORMAL	B	RIESGO ALTO
F	A	SOLTERO	A	A	PRO	D	OTROS	NORMAL	A	RIESGO ALTO
F	C	VIUDO	B	B	PLUS	A	OTROS	NORMAL	B	RIESGO BAJO
M	B	SOLTERO	A	B	PRO	B	CONYUGE	NORMAL	A	RIESGO BAJO
M	B	SOLTERO	A	C	PRO	A	HIJO(A)	PROBLEMAS POTENCIALES	A	RIESGO BAJO
M	A	DIVORCIADO	B	C	PRO	A	TITULAR	PROBLEMAS POTENCIALES	B	RIESGO BAJO
M	B	VIUDO	C	C	PRO	A	TITULAR	PROBLEMAS POTENCIALES	B	RIESGO BAJO
F	B	CASADO	C	C	CLASICO	B	TITULAR	NORMAL	C	RIESGO BAJO
F	A	DIVORCIADO	B	C	PRO	B	TITULAR	NORMAL	B	RIESGO BAJO
M	B	SOLTERO	C	C	PRO	B	TITULAR	NORMAL	A	RIESGO BAJO
M	C	DIVORCIADO	B	C	PLUS	A	TITULAR	NORMAL	C	RIESGO BAJO
M	A	CASADO	A	C	PRO	A	TITULAR	NORMAL	A	RIESGO BAJO
F	A	SOLTERO	B	B	PLUS	B	OTROS	PROBLEMAS POTENCIALES	A	RIESGO BAJO
F	C	SOLTERO	A	C	PLUS	B	TITULAR	NORMAL	A	RIESGO BAJO
F	C	DIVORCIADO	C	A	PLUS	B	TITULAR	PROBLEMAS POTENCIALES	C	RIESGO ALTO
F	A	SOLTERO	A	C	PLUS	A	TITULAR	DEFICIENTE	B	RIESGO BAJO
F	C	CASADO	B	C	CLASICO	A	CONYUGE	PROBLEMAS POTENCIALES	A	RIESGO BAJO
F	A	SOLTERO	C	C	CLASICO	A	OTROS	NORMAL	A	RIESGO BAJO
M	C	SOLTERO	B	B	PRO	B	OTROS	NORMAL	B	RIESGO BAJO
F	B	SOLTERO	A	C	CLASICO	A	HIJO(A)	PROBLEMAS POTENCIALES	B	RIESGO BAJO
M	B	CASADO	A	A	PLUS	B	TITULAR	NORMAL	A	RIESGO BAJO

Figura 24. Datos preparados para el modelo.

3.2.4. Modelado

En esta fase de la metodología se seleccionará la técnica más apropiada para cumplir los objetivos propuestos en este proyecto de minería de datos. Una vez que se tenga los datos necesarios, se aplicará la técnica para generar el modelo para luego ser evaluados.

3.2.4.1. Escoger la técnica de modelado

Para seleccionar la técnica de modelado se utilizará el software Weka el cual ofrece distintas técnicas para para realizar minería de datos. En primer lugar, el software de minería de datos cuenta con tareas las cuales son las predictivas y descriptivas. Para cumplir con el objetivo de negocio el cual es predecir el riesgo de morosidad se utilizará la tarea predictiva, del cual se ha escogido la técnica de clasificación. Dentro de esta técnica se encuentra el método árboles de decisión, la cual será utilizada en la investigación ya que se adecua más al objetivo de la investigación. Entre los algoritmos de la herramienta Weka para los árboles de decisión se seleccione el algoritmo ID3. Este proceso de selección se observa en la siguiente tabla.

Tabla 16

Selección de la técnica de modelado

HERRAMIENTA	TAREA	TÉCNICA	MÉTODO	ALGORITMO
	Predictiva	Clasificación	Árboles de decisión	ID3

3.2.4.2. Generar el Plan de prueba

En esta fase se realiza una comprobación de la calidad y validez del modelo a utilizar, el software Weka ofrece la matriz de confusión. El software Weka genera dos alternativas que son Use training set y Supplied test set.

- Use training set (Usar el conjunto de entrenamiento), es la carga de datos para el entrenamiento del modelo con todos los datos disponibles.
- Supplied test set (Conjunto de prueba suministrada), es la carga de datos con los cuales se realiza la evaluación de pruebas del modelo.

Matriz de confusión: Permite visualizar mediante una tabla de contingencia la distribución de errores cometidos por un clasificador. La matriz para el caso de dos clases se muestra de la siguiente manera:

Tabla 17

Matriz de confusión

	Clase a predecir	
Caso Real	**SI**	**NO**
SI	TP	FN
NO	FP	TN

Fuente: Santamaría W. (2010, p.44).

- **TP:** Hace referencia a true positives, es decir observaciones que el modelo clasifico correctamente para cada clase.
- **TN:** Hace referencia al total de negativos, es decir observaciones que el modelo clasifico incorrectamente para cada clase.
- **VP:** Hace referencia a la cantidad de positivos que fueron clasificados correctamente como positivos.
- **VN:** Hace referencia a la cantidad de negativos que fueron clasificados correctamente como negativos.

Para calcular la precisión del modelo, se utiliza la siguiente formula:

$$Precision = \frac{TP}{Total} \quad (7)$$

Para calcular la tase de error, se utiliza la siguiente formula:

$$Tasa\ de\ error = \frac{TN}{Total} \quad (8)$$

Para calcular la sensibilidad:

$$Sensibilidad = \frac{VP}{Total\ Positivos} \quad (9)$$

Para calcular la especificidad:

$$Especificidad = \frac{VN}{Total\ Negativos} \quad (10)$$

3.2.4.3. Construir el modelo

En esta etapa se generó el modelo que se ha elegido en los datos de entrenamiento. El software ofrece una filtración de datos mediante algoritmos con el fin de obtener los atributos de interés para obtener una mayor factibilidad.

Para la investigación se realizó el árbol de decisión con el modelo matemático para representar como se crea los árboles de decisión. Para realizar el modelo matemático del algoritmo ID3 se ha seleccionado solo algunas variables del modelo real puesto que sería muy extenso si se tomará todas las variables y todos los datos de los clientes. Por lo cual se va a ejemplificar el modelo de la siguiente tabla de datos.

COD_AFI	SEXO	TIEMPO_TRAB	INGRESO_MEN	PROGRAMA	RIESGO
1	F	C	C	PLUS	RIESGO BAJO
2	F	A	C	PLUS	RIESGO BAJO
3	F	C	B	CLASICO	RIESGO BAJO
4	F	B	A	CLASICO	RIESGO ALTO
5	F	B	B	PRO	RIESGO BAJO
6	M	A	B	PLUS	RIESGO ALTO
7	F	A	A	CLASICO	RIESGO BAJO
8	F	A	C	PRO	RIESGO ALTO
9	F	A	B	CLASICO	RIESGO BAJO
10	M	A	B	PLUS	RIESGO ALTO
11	F	C	A	PLUS	RIESGO ALTO
12	M	C	B	PLUS	RIESGO BAJO
13	F	A	B	PLUS	RIESGO BAJO
14	M	A	B	CLASICO	RIESGO BAJO
15	F	A	C	CLASICO	RIESGO BAJO
16	M	A	C	CLASICO	RIESGO BAJO
17	M	A	B	PRO	RIESGO BAJO
18	M	A	A	CLASICO	RIESGO ALTO
19	M	B	A	PLUS	RIESGO ALTO
20	F	B	C	PLUS	RIESGO BAJO

Figura 25. Data de los clientes de la empresa.

Una vez que se ha obtenido los datos con las variables para la predicción, se empieza a utilizar la fórmula de la entropía, es decir seleccionar el mejor atributo.

Primero se realiza la siguiente fórmula para calcular la información de la clase general de la variable predictiva:

- Riesgo bajo = 13/20
- Riesgo alto = 7/20

$$I\left(\frac{13}{20},\frac{7}{20}\right) = -\frac{13}{20}.\log_2\frac{13}{20} - \frac{7}{20}.\log_2\frac{7}{20}$$

$$I\left(\frac{13}{20},\frac{7}{20}\right) = 0.403 + 0.529 = 0.932$$

Segundo se realizó la misma fórmula para la calcular la información pero ahora de cada variable o atributo.

- Sexo = F

$$I\left(\frac{9}{12},\frac{3}{12}\right) = -\frac{9}{12}.\log_2\frac{9}{12} - \frac{3}{12}.\log_2\frac{3}{12} = 0.811$$

- Sexo = M

$$I\left(\frac{4}{8},\frac{4}{8}\right) = -\frac{4}{8}.\log_2\frac{4}{8} - \frac{4}{8}.\log_2\frac{4}{8} = 1$$

- Tiempo_trab = A

$$I\left(\frac{8}{12},\frac{4}{12}\right) = -\frac{8}{12}.\log_2\frac{8}{12} - \frac{4}{12}.\log_2\frac{4}{12} = 0.918$$

- Tiempo_trab = B

$$I\left(\frac{2}{4},\frac{2}{4}\right) = -\frac{2}{4}.\log_2\frac{2}{4} - \frac{2}{4}.\log_2\frac{2}{4} = 1$$

- Tiempo_trab = C

$$I\left(\frac{3}{4},\frac{1}{4}\right) = -\frac{3}{4}.\log_2\frac{3}{4} - \frac{1}{4}.\log_2\frac{1}{4} = 0.311 + 0.5 = 0.811$$

- Ingreso_Men = A

$$I\left(\frac{1}{5},\frac{4}{5}\right) = -\frac{1}{5}.\log_2\frac{1}{5} - \frac{4}{5}.\log_2\frac{4}{5} = 0.464 + 0.256 = 0.72$$

- Ingreso_Men = B

$$I\left(\frac{7}{9},\frac{2}{9}\right) = -\frac{7}{9}.\log_2\frac{7}{9} - \frac{2}{9}.\log_2\frac{2}{9} = 0.282 + 0.481 = 0.763$$

- Ingreso_Men = C

$$I\left(\frac{5}{6},\frac{1}{6}\right) = -\frac{5}{6}.\log_2\frac{5}{6} - \frac{1}{6}.\log_2\frac{1}{6} = 0.219 + 0.429 = 0.648$$

- Programa = CLÁSICO

$$I\left(\frac{6}{8},\frac{2}{8}\right) = -\frac{6}{8}.\log_2\frac{6}{8} - \frac{2}{8}.\log_2\frac{2}{8} = 0.311 + 0.5 = 0.811$$

- Programa = PRO

$$I\left(\frac{2}{3},\frac{1}{3}\right) = -\frac{2}{3}.\log_2\frac{2}{3} - \frac{1}{3}.\log_2\frac{1}{3} = 0.390 + 0.528 = 0.918$$

- Programa = PLUS

$$I\left(\frac{5}{9},\frac{4}{9}\right) = -\frac{5}{9}.\log_2\frac{5}{9} - \frac{4}{9}.\log_2\frac{4}{9} = 0.471 + 0.519 = 0.99$$

Ganancia de la información

Como ya se obtuvo la entropía de cada variable se procede a obtener la ganancia de la información utilizando la siguiente formula:

Tabla 18

Ganancia de información de variables

Variables	Ganancia de información
SEXO	0.932-[12/20(0.811) + 8/20(1)] =0.045
TIEMPO_TRAB	0.932-[12/20(0.918) + 4/20(1) + 4/20(0.811)] = 0.02
INGRESO_MEN	0.932-[5/20(0.72)+9/20(0.763)+6/20(0.648)]= 0.215
PROGRAMA	0.932-[8/20(0.811) + 3/20(0.918) + 9/20(0.99)] = 0.026

Una vez que se ha obtenido la ganancia de información de cada variable, se empieza a construir el árbol, la variable que tiene mayor ganancia de información es decir es el mejor atributo, pasa como nodo raíz del árbol, en este caso la variable INGRESO_MEN. El árbol queda de la siguiente manera como se muestra la figura 26.

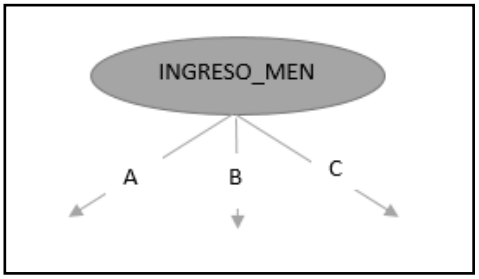


Figura 26. Primera parte del árbol de decisión.

Como se observa en la base de datos ninguna de las tres categorías del atributo INGRESO_MEN tiene una sola respuesta, por lo que se necesita realizar el mismo procedimiento para seguir construyendo el árbol pero sin contar el atributo INGRESO_MEN ya que fue tomado como raíz del árbol.

- Riesgo bajo = 1/5
- Riesgo alto = 4/5

$$I\left(\frac{1}{5},\frac{4}{5}\right) = -\frac{1}{5}.\log_2\frac{1}{5} - \frac{4}{5}.\log_2\frac{4}{5} = 0.464 + 0.256 = 0.72$$

Como se obtuvo la entropía de la variable predictiva solo del atributo INGRESO_MEN “A”, se realiza lo mismo para las demás variables.

Tabla 19

Cálculo de la entropía

Variable	Valores	Entropía
Sexo	F	$I(C) = -\frac{1}{3}.\log_2\frac{1}{3} - \frac{2}{3}.\log_2\frac{2}{3} = 0.528 + 0.39 = 0.918$
	M	$I(C) = -\frac{0}{2}.\log_2\frac{0}{2} - \frac{2}{2}.\log_2\frac{2}{2} = 0 + 0 = 0$
Tiempo de trabajo	A	$I(C) = -\frac{1}{2}.\log_2\frac{1}{2} - \frac{1}{2}.\log_2\frac{1}{2} = 0.5 + 0.5 = 1$
	B	$I(C) = -\frac{0}{2}.\log_2\frac{0}{2} - \frac{2}{2}.\log_2\frac{2}{2} = 0$
	C	$I(C) = -\frac{0}{1}.\log_2\frac{0}{1} - \frac{1}{1}.\log_2\frac{1}{1} = 0$

Programa	CLÁSICO	$I(C) = -\frac{1}{3}.\log_2\frac{1}{3} - \frac{2}{3}.\log_2\frac{2}{3} = 0.528 + 0.39 = 0.918$
	PLUS	$I(C) = -\frac{0}{2}.\log_2\frac{0}{2} - \frac{2}{2}.\log_2\frac{2}{2} = 0$

Obtenido la entropía de los atributos se pasa a generar la ganancia de la información con los valores de las variables.

Tabla 20

Ganancia de información de variables

Variables	Ganancia de información
SEXO	0.72-[3/5(0.918)+2/5(0)] = 0.169
TIEMPO_TRAB	0.72-[2/5(1)+2/5(0)+1/5 (0)]= 0.72
PROGRAMA	0.72-[3/5(0.918)+2/5(0)] = 0.169

Como se puede observar en la tabla 19 la variable que cuenta con mayor ganancia de información es “Tiempo_trab”, lo cual pasa como siguiente parte del árbol, como se muestra en la siguiente figura.

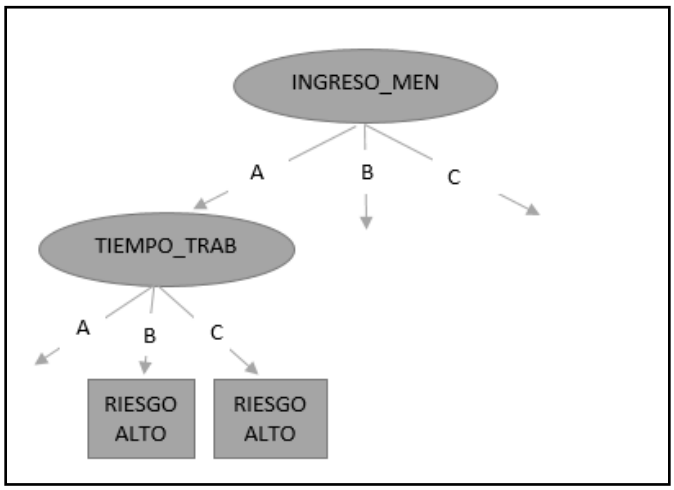


Figura 27. Segunda parte del árbol de decisión.

A continuación, se realiza el mismo procedimiento para el INGRESO_MEN=B, para lo cual se evalúa la información de la clase.

- Riesgo bajo = 7/9
- Riesgo alto = 2/9

$$I\left(\frac{7}{9},\frac{2}{9}\right) = -\frac{7}{9}.\log_2\frac{7}{9} - \frac{2}{9}.\log_2\frac{2}{9} = 0.282 + 0.481 = 0.763$$

Se obtuvo la información de la clase, luego se realizó la entropía de las variables que se necesitan para completar el árbol.

Tabla 21

Entropía de variables

Variables	Valores	Entropía
Sexo	F	$I(C) = 0 + 0 = 0$
	M	$I(C) = 0.441 + 0.528 = 0.969$
Tiempo de trabajo	A	$I(C) = 0.39 + 0.528 = 0.918$
	B	$I(C) = 0 + 0 = 0$
	C	$I(C) = 0 + 0 = 0$
Programa	CLASICO	$I(C) = 0 + 0 = 0$
	PRO	$I(C) = 0 + 0 = 0$
	PLUS	$I(C) = 0.5 + 0.5 = 1$

Se calcula la ganancia de información de las variables:

Tabla 22

Tercera parte de ganancia de información

Variables	Ganancia de información
SEXO	0.763-[4/9(0)+5/9(0.969)]= 0.224
TIEMPO DE TRABAJO	0.763-[6/9(0.918)+1/9(0)+2/9(0)]=0.151
PROGRAMA	0.763-[3/9(0)+2/9(0)+4/9(1)]=0.318

Como se observa en la tabla 21 la variable que tiene mayor ganancia de información es la variable “Programa”, la cual pasaría como nodo de la siguiente parte del árbol, como se muestra en la siguiente figura.

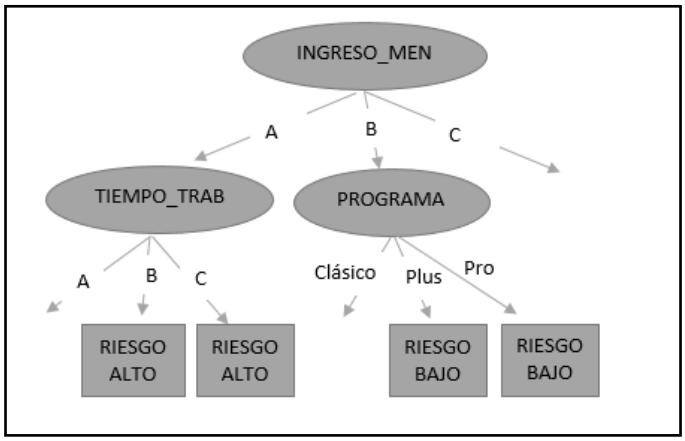


Figura 28. Tercera parte del árbol de decisión.

La variable que faltaría seria “Ingreso_men=C”, para lo cual se realiza el mismo procedimiento.

- Riesgo bajo = 5/6
- Riesgo alto = 1/6

$$I\left(\frac{5}{6},\frac{1}{6}\right) = -\frac{5}{6}.\log_2\frac{5}{6} - \frac{1}{6}.\log_2\frac{1}{6} = 0.219 + 0.429 = 0.648$$

Tabla 23

Entropía de las variables

Variables	Valores	Entropía
Sexo	F	$I(C) = 0.256 + 0.464 = 0.720$
	M	$I(C) = 0 + 0 = 0$
Tiempo de trabajo	A	$I(C) = 0.311 + 0.5 = 0.811$
	B	$I(C) = 0 + 0 = 0$
	C	$I(C) = 0 + 0 = 0$
Programa	CLASICO	$I(C) = 0 + 0 = 0$
	PRO	$I(C) = 0 + 0 = 0$
	PLUS	$I(C) = 0 + 0 = 0$

Tabla 24

Ganancia de información

Variables	Ganancia de información
SEXO	0.648-[5/6(0.720)+1/6(0)]= 0.048
TIEMPO DE TRABAJO	0.648-[4/6(0.811)+1/6(0)+1/6(0)]=0.107
PROGRAMA	0.648-[2/6(0)+1/6(0)+3/6(0)]=0.648

Como se observa en la tabla anterior la variable con mayor ganancia es Programa, la cual pasa como nodo del atributo “Ingreso_men=C”. El árbol quedaría de la siguiente manera:

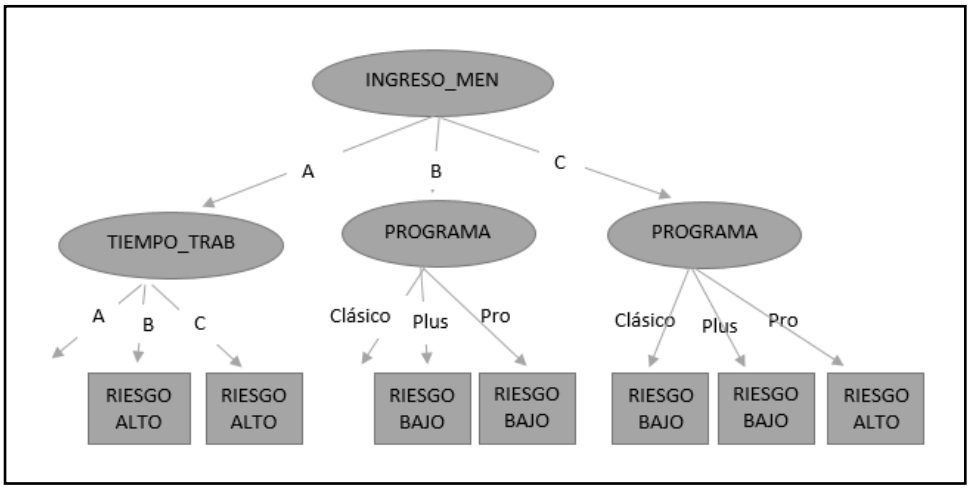


Figura 29. Modelo de árbol de decisión completo.

Como se observa en la figura 30, el árbol se ha construido hasta que el nodo raíz sea la variable predictiva, esto se ha realizado mediante el modelo matemático, con el fin de conocer cómo se construye el algoritmo ID3 mediante esta forma.

Una vez creado este modelo de manera matemática, se va a utilizar la herramienta Weka para todos los datos.

- **Modelo:** El software Weka señala que los datos de entrenamiento corresponden al 70% y el 30% restante a los datos de prueba.
- **Descripción del modelo:** Se describen los resultados que devuelve el modelo en la evaluación.

La carga de los datos de entrenamiento en weka se visualiza de la siguiente manera:

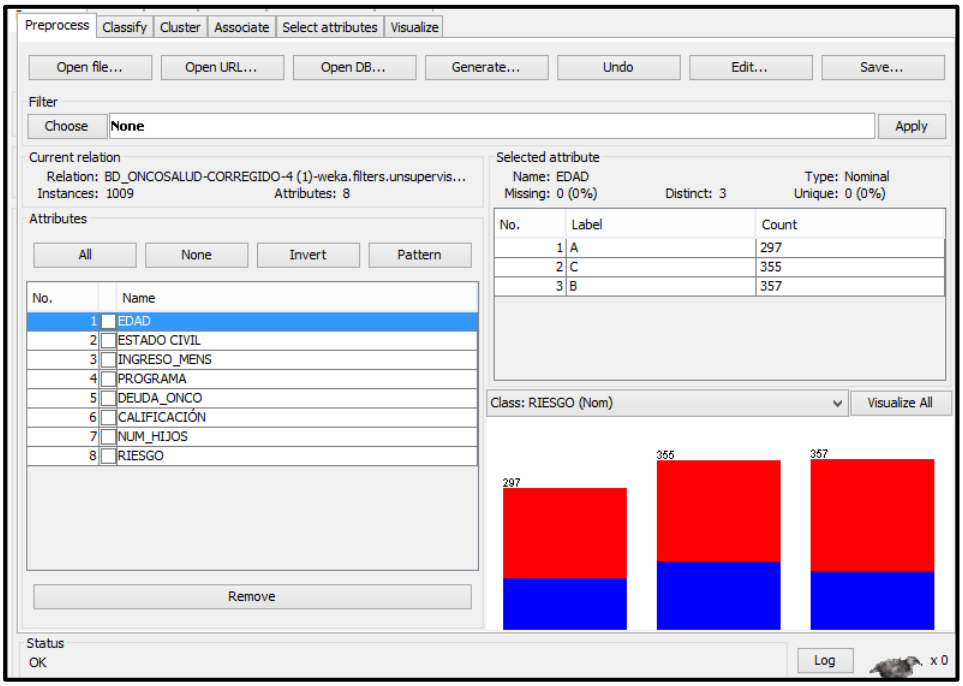


Figura 30. Datos en Weka.

Una vez que se ha ingresado los datos en la herramienta Weka se procede a seleccionar el algoritmo de árboles de decisión que se ha elegido, en este caso es el ID3 y se selecciona el label “RIESGO”.

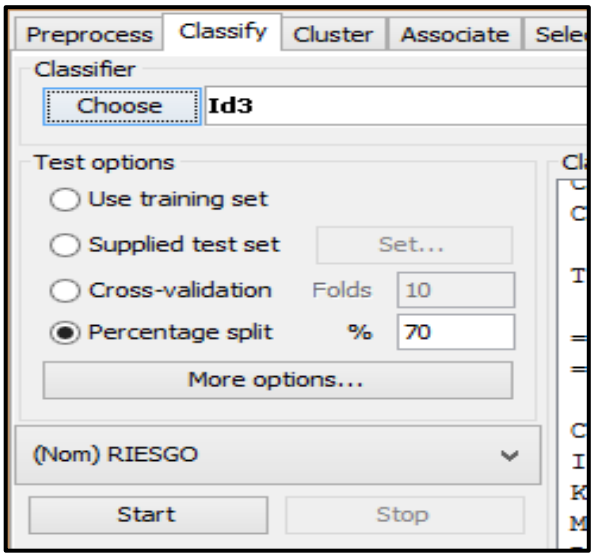


Figura 31. Selección del algoritmo ID3 en Weka.

La siguiente figura muestra los resultados que muestra el algoritmo ID3 de árboles de decisión solo utilizando los datos de entrenamiento.

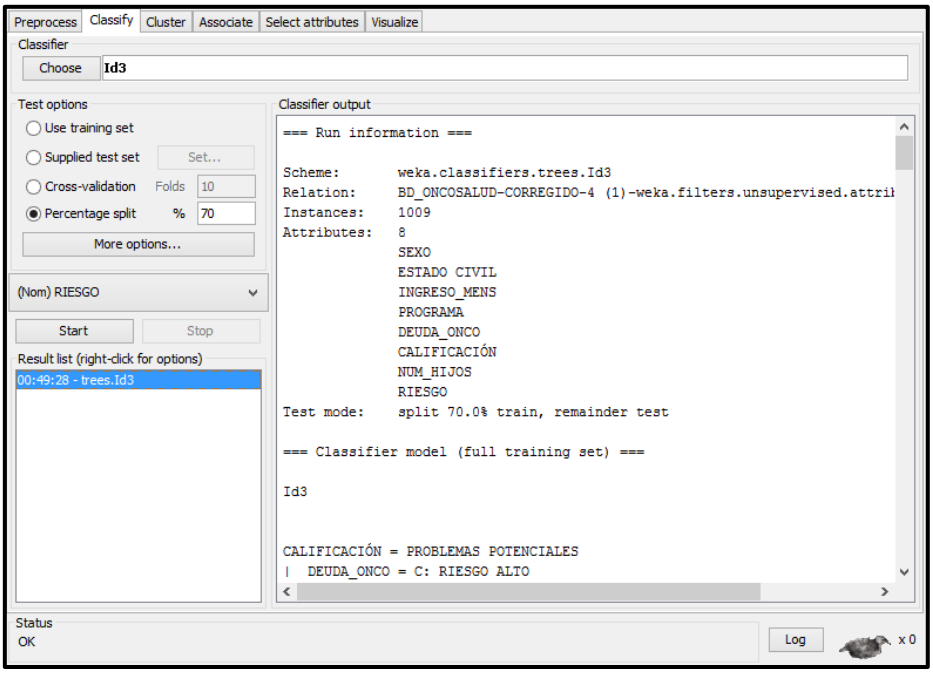


Figura 32. Resultados del algoritmo ID3.

En este resultado muestra el algoritmo, es decir reglas a seguir para predecir si un cliente podría ser moroso o no. Ver Anexo VIII.

Descripción del modelo

El modelo ID3 ha devuelto los siguientes resultados:

- Confianza predictiva para el algoritmo ID3 ha dado un valor de 88.1%.
- Error absoluto medio para el algoritmo ID3 ha dado un valor de 9.2%.
- Error cuadrático medio para el algoritmo ID3 ha dado un valor de 30%.

3.2.4.4. Evaluar el modelo

En esta etapa se realiza una evaluación a los objetivos de la minería de datos que se han propuesto en la investigación. Una buena forma de evaluar la efectividad del modelo son los indicadores estadísticos que ofrece el software Weka que han sido señalados en el plan de prueba. Como son la confianza predictiva, el error cuadrático medio y error absoluto medio.

Para el modelo se puede decir que la confianza predictiva que ha mostrado tiene un valor de 88.1%, lo cual da un buen resultado para el objetivo propuesto de minería de datos.

La matriz de confusión generada por el árbol de decisión y el algoritmo ID3 es la siguiente:

```
=== Confusion Matrix ===

   a   b   <-- classified as
  98  14 |   a = RIESGO ALTO
  14 169 |   b = RIESGO BAJO
```

Figura 33. Matriz de confusión del modelo.

La precisión está dado por:

$$Precision = \frac{98 + 169}{98 + 14 + 14 + 169} = 88.1$$

La tasa de error está dado por:

$$Error = \frac{14 + 14}{98 + 14 + 14 + 169} = 0.09$$

La sensibilidad está dado por:

$$Sensibilidad = \frac{98}{98 + 14} = 0.87$$

La especificidad está dado por

$$Especificidad = \frac{169}{169 + 14} = 0.92$$

En la matriz de confusión se muestra que el grado de predicción es bastante bueno puesto que tiene un 88% de precisión del modelo de árboles de decisión. En cuanto a la tasa de error ha dado un 9%. También se observa otras medidas como la sensibilidad, que quiere decir que tan bueno predice a los clientes con un riesgo alto ha dado un 87% y la especificidad, que tan bien predice el modelo a los clientes con un riesgo bajo ha dado un 92%.

3.2.5. Evaluación

En esta fase de la metodología se realizará una evaluación a los objetivos del negocio establecidos al inicio de la investigación. Además, se revisa el proceso si en caso se ha cometido un error para corregirlo.

3.2.5.1. Evaluar resultado

Para realizar este proceso se establece que para medir el resultado que otorga la herramienta de software Weka se debe tener en cuenta los indicadores estadísticos que nos muestra esta herramienta como es la precisión y matriz de confusión, el cual se ha realizado en la fase del modelado.

Al realizar el modelo se ha llegado a la conclusión que el modelo es aceptable ya que su grado de precisión supera el 88% como se observa en la tabla 24.

Tabla 25

Resultado de la matriz de confusión del modelo desarrollado

Matriz de confusión	Árbol de decisión ID3
Precisión	88%
Tasa de error	12%
Sensibilidad	87%
Especificidad	92%

3.2.5.2. Revisar el proceso

En este proceso se establece que los datos obtenidos de la muestra de los clientes fueron extraídos de la misma base de datos de la empresa lo cual se puede garantizar que los datos fueron correctos e íntegros.

3.2.5.3. Determinar los próximos pasos

Dado que los resultados han sido favorables a los objetivos del negocio, el siguiente paso es realizar la fase de implantación del proyecto.

3.2.6. Implantación

En esta última fase de la metodología CRISP-DM se documenta y presenta los resultados al usuario con el fin de que tenga una fácil comprensión y un mayor conocimiento.

3.2.6.1. Planear la implantación

Para realizar la implantación se debería tener un mayor acceso a la base de datos de la empresa Oncosalud, puesto que algunos atributos han sido extraídos de fuentes externas para la realización de esta investigación, lo cual podría afectar en la predicción del modelo. Asimismo, será necesario introducir los atributos en una sola base de datos, ya sea los datos anteriores como los actuales. Por consiguientes, una vez obtenido todos los datos de la empresa tomaría más tiempo realizar este proceso ya que la cantidad de datos que maneja la empresa es mayor.

Para mostrar los resultados de la minería de datos se creó un sistema web, el cual a continuación se explicará cómo se realizó la integración de weka con JSP, que es una tecnología orientado a la creación de páginas web dinámicas basado en HTML usando el lenguaje de programación JAVA, teniendo como propósito mostrar el nivel de riesgo que puede tener cada cliente en base a su información o variables que se plantearon en esta investigación utilizando el algoritmo ID3 de Weka.

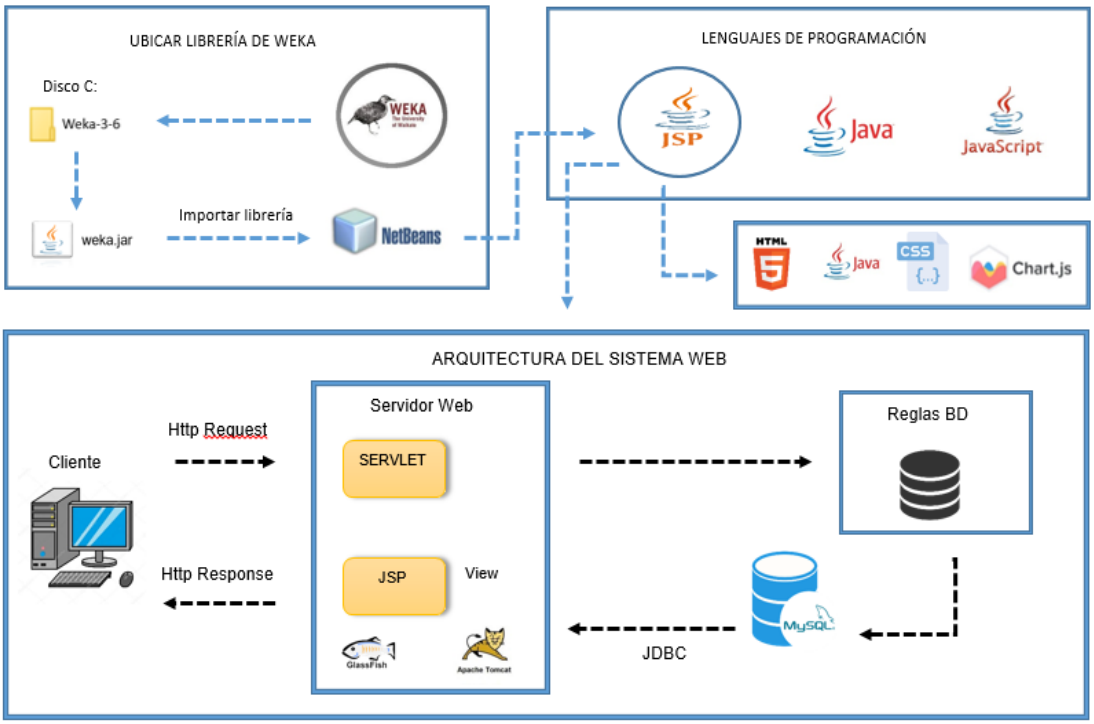


Figura 34. Modelo de la arquitectura del sistema.

Etapas para integrar Weka con Java:

ETAPA 1: Ubicar librería de Weka

Como primer paso para poder realizar la integración de Weka con JSP se tuvo que realizar la búsqueda de las librerías de Weka para poder ser utilizadas en el desarrollo del sistema web. Esta librería está comprendida por un JAR, teniendo como nombre weka.jar que se encuentra ubicado en la misma carpeta de Weka en el disco C al momento de haber realizado su instalación.

- Weka.jar: Contiene librerías de la técnica de minería de datos como por ejemplo clustering, clasificación, asociación utilizadas en el software Weka.

Luego de haber ubicado el weka.jar se realiza la importación de este JAR al entorno Netbeans para dar inicio la programación del sistema web.

ETAPA 2: Lenguajes de programación

Como segundo paso se debe elegir bajo que entorno quieres trabajar con el fin de poder contrastar si se puede integrar con Weka y no generar problemas de desarrollo más adelante, en este caso se eligió trabajar con JSP ya que esta tecnología trabaja bajo el entorno java y puede integrarse con el software Weka de minería de datos.

Es importante tomar en cuenta respecto al llamado de la base de datos que vamos a utilizar al momento de realizar la programación con las librerías de Weka, ya que trabaja con bases de datos bajo un formato ARFF de lo contrario al momento de llamar a la base de datos con otro tipo de formato ya sea CSV que también soporta Weka generará un error.

- ARFF: Es un formato propio que utiliza Weka en sus bases de datos.

ETAPA 3: Arquitectura del sistema web

Como tercer paso debemos conocer la arquitectura de tu sistema para saber lo que se está desarrollando y tener claro lo que se está avanzando. A continuación se muestra como interactúa cada uno de los componentes de la arquitectura del aplicativo web al momento de realizar la predicción.

- PASO 1: En la arquitectura del sistema se visualiza un ejemplo desde que una persona entra al sistema y realiza una petición (Request) para poder predecir si el cliente podría tener un riesgo alto, medio y bajo.

- PASO 2: Esta solicitud del cliente es tomada por los servlet para devolver una respuesta, pero antes de devolver una respuesta estos campos ingresados por el cliente se dirige a las reglas propuestas por el modelo y realiza una comparación con cada una de ellas para poder realizar la predicción, el cual se almacenará la información ingresada en la base de datos MySql.

- PASO 3: Una vez almacenado la información en la base de datos se muestra la información solicitada por el cliente con los JSP ya que su función principal es mostrar de manera gráfica la página

ejecutándose en un servidor que puede ser APACHE TOMCAT o GLASSFISH.

3.2.6.3. Implementación del sistema Web de predicción

INTERFAZ DE LOGEO

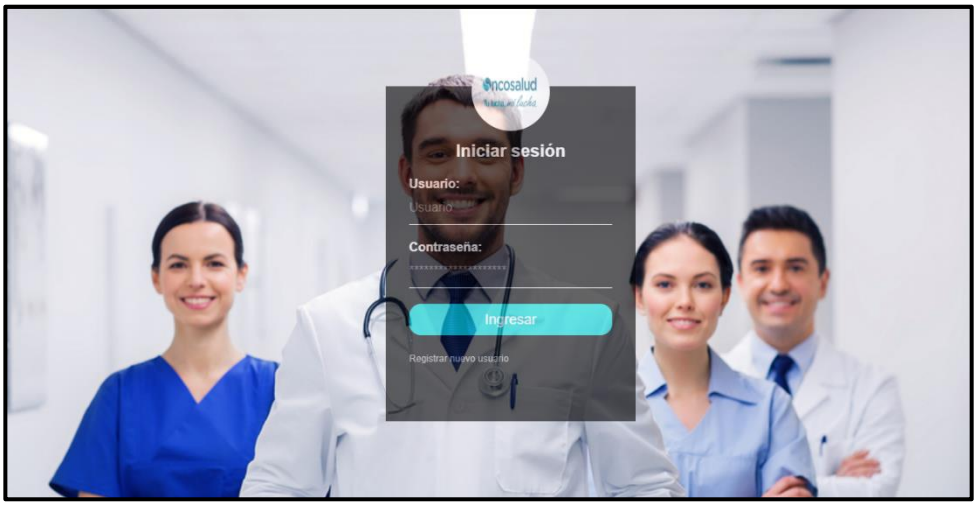


Figura 35. Pantalla de inicio de sesión del sistema.

DESCRIPCION DEL MENÚ

1. **Menú Datos históricos:** Se detalla la información histórica de los clientes que se han ingresado.

Figura 36. Listado de clientes en el sistema.

2. **Menú Algoritmo:** Se visualización del Pseudocódigo de la base de datos ejecutado con el algoritmo ID3 de weka.

Figura 37. Resultado del algoritmo ID3 generado por el sistema.

3. **Menú Cliente:** Formulario para ingresar información del cliente para realizar la predicción del riesgo de morosidad.

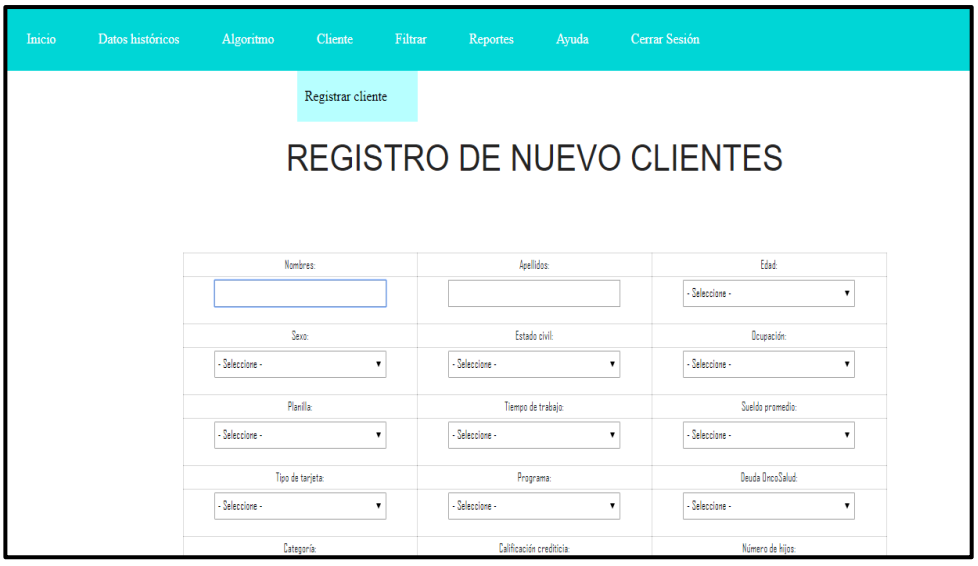


Figura 38. Registro de nuevos clientes del sistema.

4. **Menú Filtrar:** Opción para realizar la búsqueda de clientes ya sea por rango de edades o por programa.

Figura 39. Filtración de datos del sistema.

5. **Menú Reportes:** Opción de reportes para visualizar reportes de los clientes según su programa.

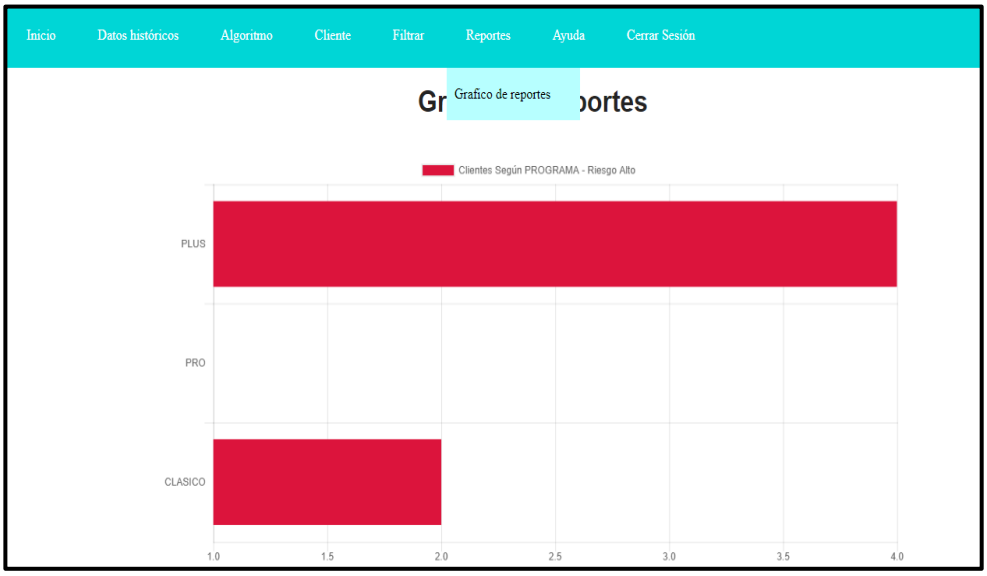


Figura 40. Gráficos de reportes del sistema.

Ejemplo del registro de un cliente para predecir su riesgo de morosidad:

- Se registra al cliente de acuerdo a la información de sus características.

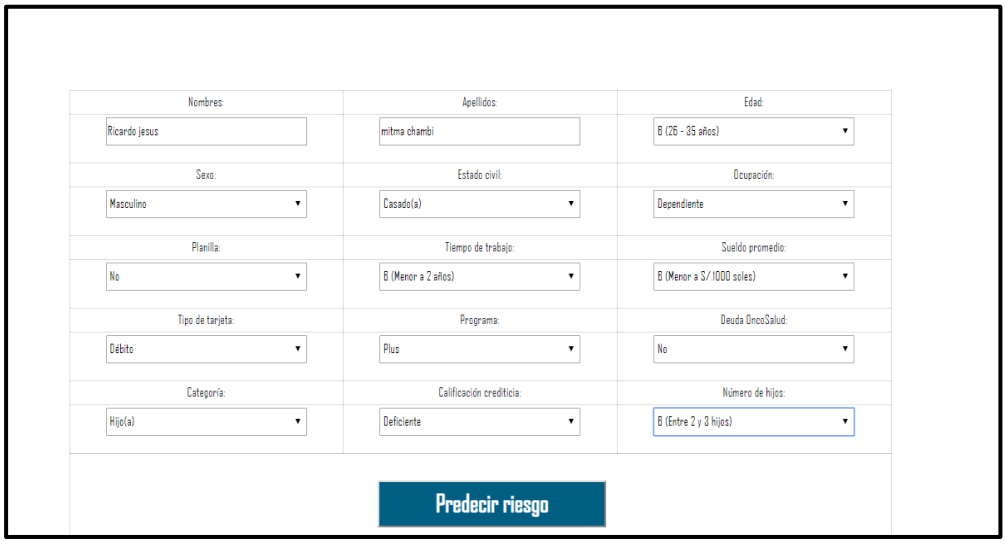


Figura 41. Registro de un cliente en el sistema.

- Se selecciona el botón predecir riesgo y nos muestra el siguiente resultado:

Figura 42. Resultado de la predicción de un cliente en el sistema.

3.2.6.4. Producir el informe final

La empresa Oncosalud vende seguros oncológicos a los clientes que deseen este servicio, los cuales el cliente debe pagar de manera mensual o anual a la empresa. Algunos clientes a los cuales se les vende este seguro tienden a ser morosos, lo cual tiende a ser una pérdida económica para la empresa. La empresa no cuenta con herramientas que les faciliten conocer que clientes podrían caer en morosidad al momento de vender un seguro. Con este proyecto de minería de datos se busca predecir el riesgo de morosidad de los clientes con el fin de tomar mejores decisiones sobre el seguro que se le desea vender.

Para realizar este proyecto se utilizó la metodología CRISP-DM ya que es una de las más utilizadas para este tipo de proyectos. La metodología da una serie de fases a seguir para alcanzar los objetivos propuestos de una manera ordenada. En el primer paso se realiza la preparación de los datos a utilizar, la cual ha tomado un mayor tiempo que otras fases, luego la evaluación de estos, para luego seleccionar un modelo y aplicarlos, por ultimo analizar los resultados para dar respuesta a los objetivos propuestos.

Se aplicó una técnica de modelado, la cual fue el algoritmo ID3 a los datos seleccionados, lo cual fue analizado en las fases de la metodología. Se obtuvo como resultado que el algoritmo ha dado una confianza aceptable lo cual ha cumplido con el objetivo del proyecto.

3.2.6.5. Revisar el proyecto

En este último proceso de la metodología CRISP-DM se analizó los procesos realizados, los que se han hecho correctamente y los que no. Asimismo, posibles mejoras para proyectos a futuro con el fin de obtener mejores resultados.

Como se ha mencionado anteriormente la fase de recolección de dato fue lo más trabajoso ya que se extrajo de diferentes fuentes para poder completar la base de datos. Si se hubiera tenido todos los datos en un solo repositorio, el resultado sería más eficaz y confiable en el modelo de minería de datos realizado en este proyecto.

CAPÍTULO IV

ANÁLISIS DE RESULTADOS Y CONTRASTACIÓN DE LA HIPÓTESIS

4.1 POBLACIÓN Y MUESTRA

4.1.1 Población

La población de estudio está constituida por los clientes de la empresa Oncosalud. Se consideró un total de 85 000 registro de clientes.

4.1.2 Muestra

Para el desarrollo de la investigación se tiene dos clases de muestras por un lado se encuentran los clientes y por otro los empleados del departamento de cobranza de la empresa Oncosalud.

Se cuenta con 1009 registros de clientes entre los meses de junio, julio y agosto del 2018, los cuales han sido seleccionados aleatoriamente.

A su vez también se les aplicará una encuesta a los 6 empleados del departamento de cobranza.

$$n = \frac{N\sigma^2 Z^2}{e^2(N-1)+\sigma^2 Z^2} \qquad (11)$$

Donde:

n = El tamaño de la muestra.

N = Tamaño de la población.

$\sigma =$ Desviación estándar de la población suele utilizarse un valor constante de 0,5.

Z = Valor obtenido mediante niveles de confianza. (95% = 1,96)

e = Límite aceptable de error muestral (0,05)

$$n = \frac{85000.\,0{,}5^2.\,1{,}96^2}{0{,}05^2(85000-1)+0{,}5^2.\,1{,}96^2} = 382.43$$

La muestra estará conformada por 382 clientes de seguros oncológicos en la empresa de seguros Oncosalud.

4.2 VALIDEZ Y CONFIABILIDAD DEL INSTRUMENTO

4.2.1 Validez

La validez del instrumento se ha realizado a través de tres expertos quienes han revisado la pertinencia, relevancia y claridad recomendando su aplicabilidad.

Tabla 26

Validez de los instrumentos por expertos

	Experto 1	**Experto 2**	**Experto 3**
Nombre	Guevara Ponce Víctor Manuel	Cuya Cámara Luis	Cabanillas Carbonell Michael
Grado	Ingeniero Estadístico	Doc. Informático	Ciencias de la computación
Institución	Universidad Autónoma del Perú	Universidad Autónoma del Perú	Universidad Autónoma del Perú

4.2.2 Confiabilidad

La confiabilidad del instrumento se ha determinado a través del método Test–Retest y Alfa de Cronbach que se aplicó a la observación de la pre prueba.

Para el indicador KPI3: Nivel de dificultad

Se utilizó el instrumento Alfa de Cronbach para determinar la homogeneidad de los datos, se observó que los datos superan el 80% con lo cual se concluye que la data es buena.

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
,833	5

Figura 43. Validez del indicador Nivel de dificultad.

Para el indicador KPI4: Tiempo para predecir

Se utilizó el Test – Retest, el cual aplica el coeficiente de correlación de Pearson para determinar la estabilidad, se observó que supera el 79% por lo cual se concluye que los datos tienen una correlación alta.

Correlaciones

		TiempoPre	TiempoPost
TiempoPre	Correlación de Pearson	1	,797**
	Sig. (bilateral)		,000
	N	382	382
TiempoPost	Correlación de Pearson	,797**	1
	Sig. (bilateral)	,000	
	N	382	382

**. La correlación es significativa en el nivel 0,01 (bilateral).

Figura 44. Validez del indicador Tiempo para predecir.

4.2 ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

4.2.1 Resultados

A continuación, se muestran los valores de los indicadores de la pre-prueba y post-prueba.

Para el indicador precisión de predicción (KPI1) y error de predicción (KPI2) se tomaron 382 clientes del mes de julio del 2018 de la empresa. Para hallar estos indicadores se realizó la matriz de confusión, véase tabla 27.

Tabla 27

Matriz de confusión para la Pre prueba del KPI1 y KPI2

	RIESGO BAJO	RIESGO ALTO
RIESGO BAJO	169	58
RIESGO ALTO	75	80

A continuación, se muestran los resultados de los indicadores:

A. Indicador 1: Precisión de la predicción: KPI1

Tabla 28

Resultado de la Pre-Prueba y Post-Prueba para el KPI1

	KPI1: Precisión de predicción	
Mes	**Pre-prueba**	**Post-prueba**
Julio 2018	65.2%	88.1%

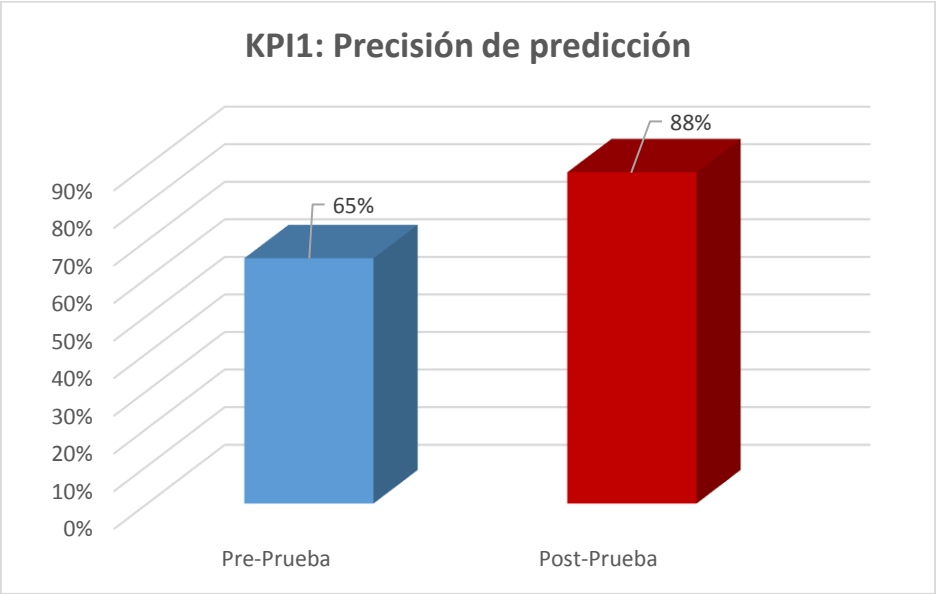


Figura 45. Precisión de la predicción de la Pre-Prueba y Post-Prueba.

Se obtuvo como resultado del indicador precisión de la predicción (KPI1) del modelo un porcentaje de 65% en la pre-prueba es decir sin el modelo de árboles de decisión; para la post-prueba se obtuvo un porcentaje de 88% utilizando el modelo de árboles de decisión. Esto indica que si hay una diferencia implementando el modelo de árbol de decisión ya que predice de una mejor manera a los clientes que podrían caer en morosidad en la empresa Oncosalud.

B. Indicador 2: Error de predicción: KPI2

Para el indicador error de predicción, los resultados de la pre-prueba y post-prueba del modelo de árboles de decisión se encuentran en la tabla 29.

Tabla 29

Resultado de la Pre-prueba y Post-Prueba para el KPI2

	I2: Error de predicción	
Mes	**Pre-prueba**	**Post-prueba**
Julio 2018	34.8%	11.9%

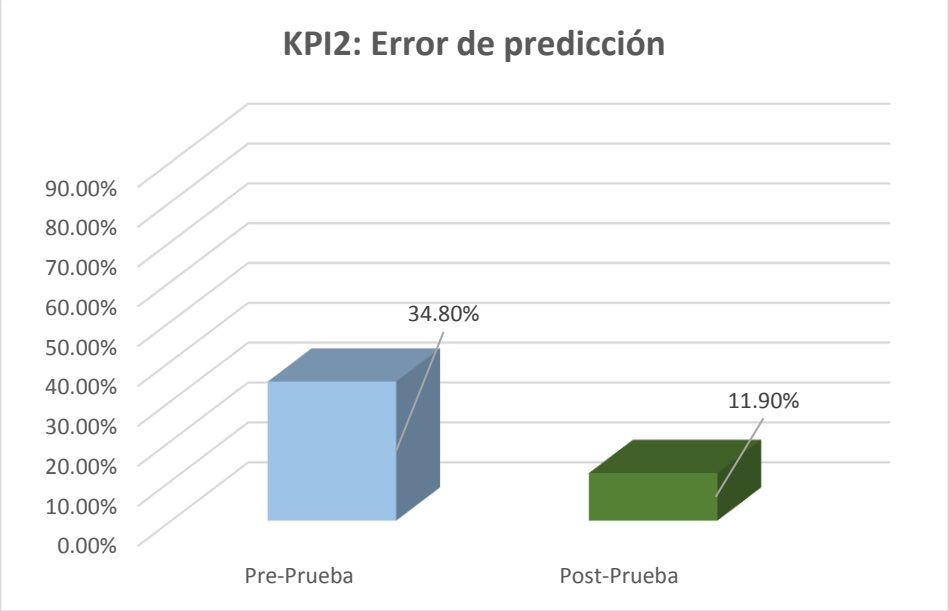


Figura 46. Error de predicción de la Pre-Prueba y Post-Prueba.

Analizando los datos para el indicador error de predicción (KPI2) del modelo, el resultado refleja que en la pre prueba se obtuvo un porcentaje de 34.8%, es decir sin el modelo de árboles de decisión, para la post prueba se obtuvo un porcentaje de 11.9% utilizando el modelo de árboles de decisión. Esto indica que utilizando el modelo de árboles de decisión disminuye el porcentaje de error del modelo para predecir el riesgo de morosidad de los clientes de la empresa.

C. Indicador 3: Nivel de dificultad: KPI3

Para el indicador nivel de dificultad se realizó un cuestionario a los empleados del departamento de cobranza antes de implementar el modelo de árboles de decisión, de los cuales se obtuvo los siguientes resultados:

Tabla 30

Resultado de la Pre-Prueba para el KPI3

	KPI3: Nivel de dificultad
N°	**Post-Prueba**
1	Difícil
2	Muy difícil
3	Difícil
4	Difícil
5	Muy difícil
6	Difícil

Tabla 31

Frecuencia de la Pre-Prueba para el KPI3

Estado	**Frecuencia**	**Porcentaje**	**Porcentaje valido**
Muy difícil	2	33%	33%
Difícil	4	67%	67%
Normal	0	0%	0%
Fácil	0	0%	0%
Muy fácil	0	0%	0%
Total	**6**	**100%**	**100%**

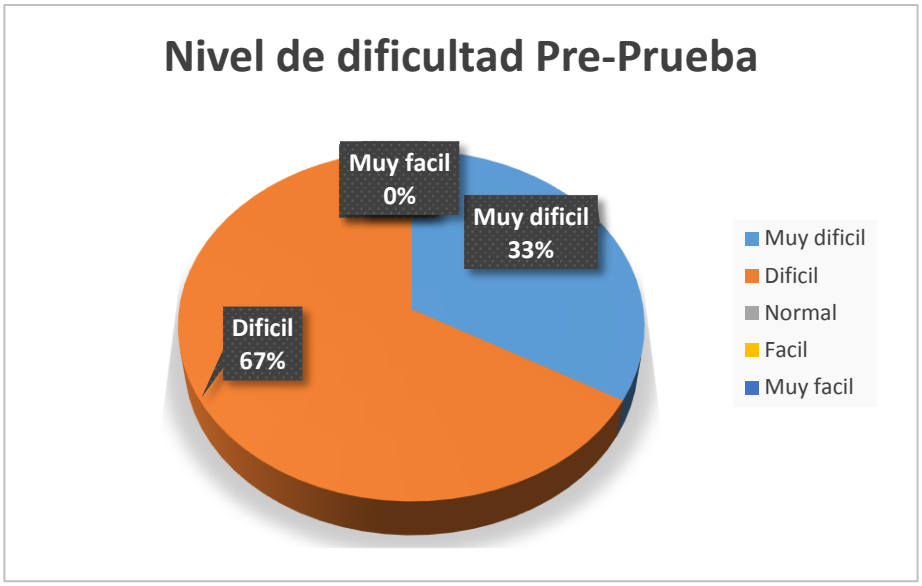


Figura 47. Nivel de dificultad Pre-prueba.

Según los datos mostrados en la figura 47 se aprecia el nivel de dificultad de los empleados antes de implementar el modelo de árboles de decisión. Según 4 empleados que representan al 67% de encuestados respondieron que el nivel de dificultad para predecir a un cliente es difícil y 2 empleados que representan al 33% respondieron muy difícil. Esto indica que el proceso de predecir el riesgo de un cliente no es tan bueno.

A continuación, se muestra los resultados del cuestionario de la post prueba implementando el modelo de árboles de decisión para la predicción de morosidad de clientes.

Tabla 32

Resultado de la Post-Prueba para el KPI3

	KPI3: Nivel de dificultad
N°	**Post-Prueba**
1	Fácil
2	Fácil
3	Muy fácil
4	Fácil
5	Muy fácil
6	Muy fácil

Tabla 33

Frecuencia de la Post-prueba para el KPI3

Estado	Frecuencia	Porcentaje	Porcentaje valido
Muy difícil	0	0%	0%
Difícil	0	0%	0%
Normal	0	0%	0%
Fácil	2	33%	33%
Muy fácil	4	67%	67%
Total	**6**	**100%**	**100%**

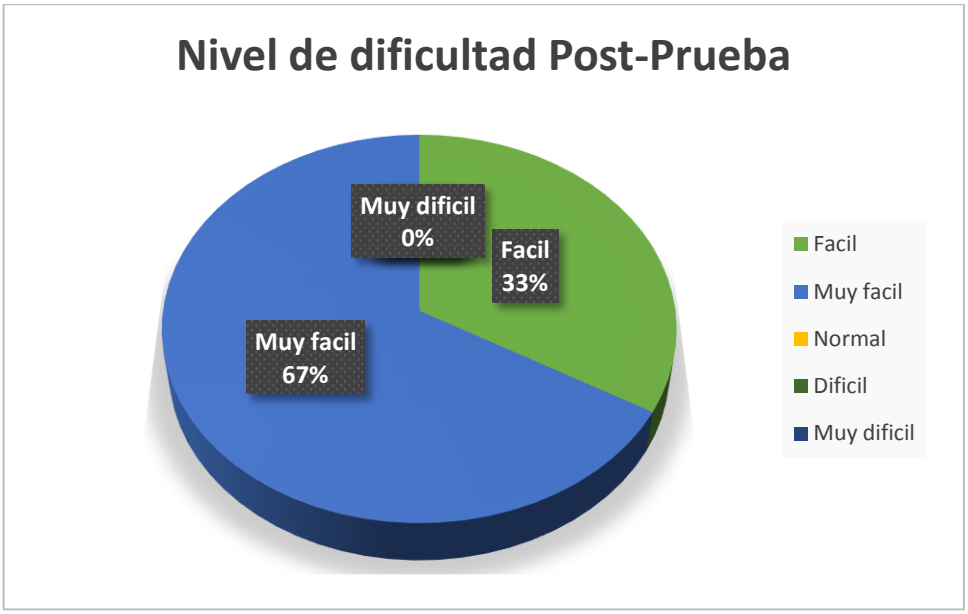


Figura 48. Nivel de dificultad para la Post-Prueba.

Según los datos mostrados en la figura 48 se aprecia el nivel de dificultad que tienen los empleados con respecto a la predicción de un cliente después de la implementación del modelo de árboles de decisión. Según 4 empleados que representan al 67% de los encuestados respondieron que el nivel de dificultad para predecir a un cliente es Muy fácil y 2 empleados que representan al 33% respondieron fácil. Esto indica que el proceso para predecir el riesgo de un cliente mejoró considerablemente con la implementación de la técnica de minería de datos.

D. Indicador 4: Tiempo para predecir: KPI4

Para el indicador Tiempo para predecir (KPI4) los resultados se encuentran en el anexo XI.

A continuación, se muestra los resultados de la pre-prueba:

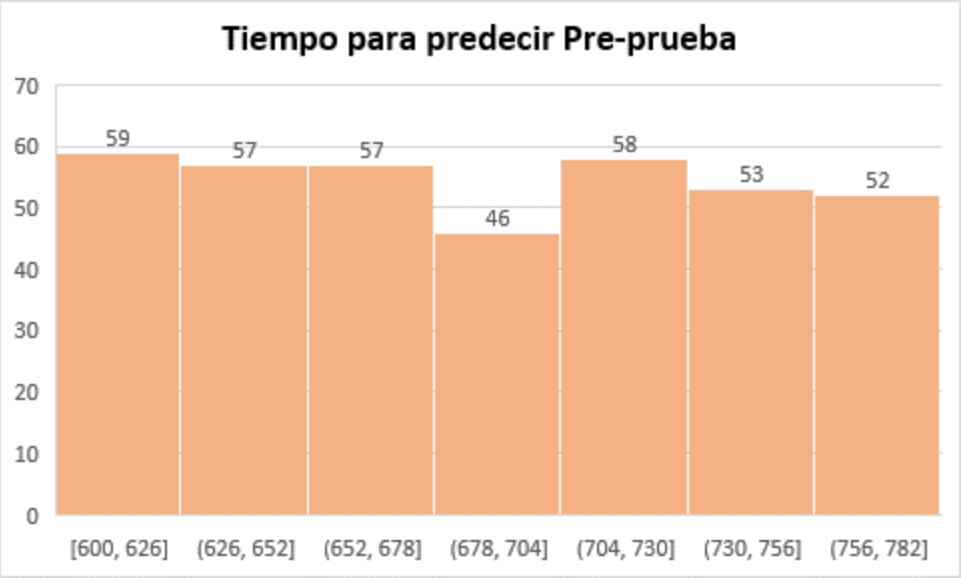


Figura 49. Resultado del tiempo para predecir (KPI4) en la Pre-Prueba.

Según los datos mostrados en la figura 49, los empleados del departamento de cobranza se demoran entre 600 a 626 segundos (10 minutos) para predecir el riesgo de morosidad de unos 59 casos de predicciones. Para 57 casos de predicciones se requirieron un tiempo de 626 a 652 segundos. Los casos de predicciones que se tardaron más corresponden a 52 casos con un tiempo comprendido entre 756 y 782 segundos, siendo unos 13 minutos para predecir el riesgo de morosidad de clientes. El tiempo promedio para predecir el riesgo de morosidad fue de 689,45 segundos sin la implementación del modelo de árboles de decisión en la pre-prueba.

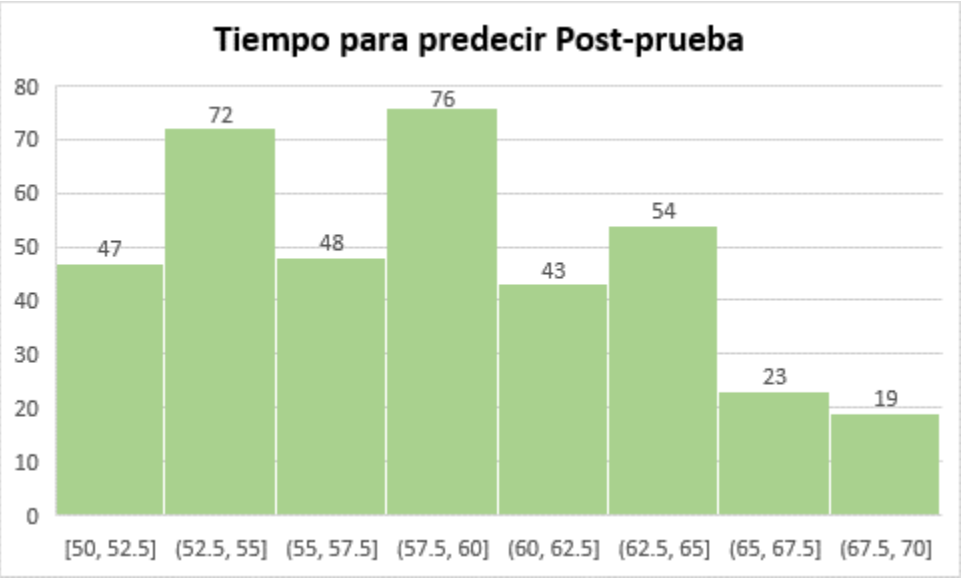


Figura 50. Resultado del tiempo para predecir (KPI4) en la Post-Prueba.

En la figura 50 se muestran los resultados del indicador tiempo para predecir de la post-prueba, es decir implementado el modelo de árboles de decisión. Se observa que el mínimo tiempo que se demoran los empleados fue entre 50 a 52,5 segundos para predecir el riesgo de morosidad de los clientes en un total de 47 casos. El máximo tiempo para predecir que se obtuvo fue de 67,5 a 70 segundos que se tardan en una predicción, con un total de 19 casos realizados. El promedio de tiempo que se demoran para la predicción con árboles de decisión fue de 58,70 segundos en la post-prueba.

A continuación, se realiza una comparación de los tiempos promedio de la pre-prueba y post-prueba.

Tabla 34

Comparación de los tiempos promedios del KPI4

Pre-prueba		**Post-prueba**		**Diferencia**	
Tiempo (seg)	**%**	**Tiempo (seg)**	**%**	**Tiempo (seg)**	**%**
689,45	100	58,70	9	630,75	91

En la tabla 34 se observa que antes de la implementación de la técnica de minería de datos basado en árboles de decisión, el tiempo de predicción de riesgo de morosidad fue de 689,45 segundos (100%), y con la implementación de la técnica de minería de datos basado en árboles de decisión, el tiempo promedio es de 58,70 segundos (9%). Lo que significa que ha reducido en 630,75 segundos (91%) con respecto a la predicción de riesgo de morosidad de los clientes.

4.3 NIVEL DE CONFIANZA Y GRADO DE SIGNIFICANCIA

Para que los datos recolectados sean evaluados se consideró los siguientes parámetros:

- El nivel de confianza será del 95%
- El nivel de significancia será del 5%

4.4 PRUEBA DE HIPÓTESIS

Para la investigación se presentan 4 indicadores:

Tabla 35

Indicadores para la Contrastación de la hipótesis

Indicador	Pre-Prueba	Post-Prueba	Comentario
I1: Precisión de la predicción	65%	88%	---
I2:Error de predicción	35%	12%	---
I3: Nivel de dificultad			indicador cualitativo
I4: Tiempo para predecir	689,45 segundos	58,70 segundos	---

- **Contrastación para la precisión de predicción**

  H_o: La aplicación de minería de datos basado en árboles de decisión no mejora la precisión de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

  H_1: La aplicación de minería de datos basado en árboles de decisión mejora la precisión de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

H_0: $P_A \leq P_{SA}$

H_1: $P_A > P_{SA}$

Donde:

- P_{SA} = Proporción de la precisión de predicción en la Pre-Prueba.
- P_A = Proporción de la precisión de predicción en la Post-Prueba.
- n_1 = Tamaño de la muestra

El método estadístico que se utilizó para comprobar la hipótesis fue la diferencia de proporciones ya que las respuestas que se obtuvieron del indicador se encuentran en porcentaje.

La diferencia de proporciones se calcula a través de la siguiente formula:

$$Z_C = \frac{P_A - P_{SA}}{\sqrt{\frac{P_A(1-P_A)}{n1} + \frac{P_{SA}(1-P_{SA})}{n2}}} \quad (12)$$

Se reemplazan los resultados de la pre-prueba y post-prueba en la fórmula:

$$Z_{C=} \frac{0.88 - 0.65}{\sqrt{\frac{0.88(1-0.88)}{382} + \frac{0.65(1-0.65)}{382}}} = 4{,}3662$$

Región crítica

Se basó en el teorema de límite central (Z), este teorema señala que si la muestra es mayor a 30 (n>30), sea cual sea la distribución de la media muestral, siempre será una distribución normal.

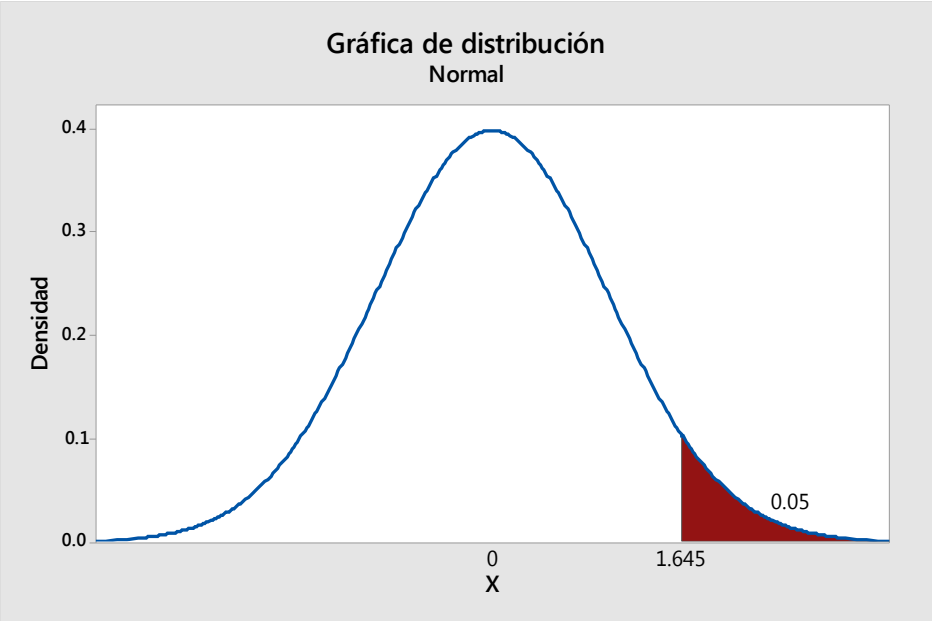


Figura 51. Gráfica de distribución KPI1.

Decisión:

Puesto que el valor de $Zc > Z_{1-\propto,n}$ (4,3662 > 1.645) entonces se rechaza la hipótesis nula (H_0) y se acepta la hipótesis alterna (H_1).

Conclusión:

Se ha demostrado que la proporción de la precisión de árboles de decisión es mayor que la proporción de la predicción sin árboles de decisión con un 95% de confianza. Efectivamente la aplicación de minería de datos basado en árboles de decisión si mejora significativamente la precisión de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

- **Contrastación para el nivel de dificultad**

  H_0: La aplicación de minería de datos basado en árboles de decisión no reduce la dificultad de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

  H_1: La aplicación de minería de datos basado en árboles de decisión reduce la dificultad de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

H_0: $\bar{X}1 \leq \bar{X}2$

H_1: $\bar{X}1 > \bar{X}2$

Donde:

- $\bar{X}1$ = Nivel de dificultad para predecir el riesgo de morosidad de la empresa de seguros aplicando la forma tradicional.
- $\bar{X}2$ = Nivel de dificultad para predecir el riesgo de morosidad de la empresa de seguros aplicando minería de datos basado en árboles de decisión.

Estadístico de prueba

Para la elección de la prueba estadística se utilizó la prueba t student ya que el tamaño de la muestra es menor que 30 (n<30).

Cálculo del valor de t student estadístico

Los resultados de la prueba t student obtenidos se muestran en las tablas 36 y 37.

Tabla 36

Estadística de muestras relacionadas del indicador nivel de dificultad KPI3

Tipo		N	Media	Desviación estándar	Media de error estándar
Dificultad	**Pre-prueba**	6	8,000	2,000	0,816
	Post-prueba	6	23,00	1,673	0,683

Tabla 37

Resultados de la prueba t student para el KPI3

Prueba t student para muestras relacionadas						
Diferencia de medias	**Desv. Error promedio**	**95% de intervalo de confianza de la diferencia**		**t**	**gl**	**Sig. (bilateral)**
		Inferior	**Superior**			
-15,000	1,183	-18,041	-11,958	-12,667	5	,000

Región crítica

El valor crítico de la prueba t student con 5 grados de libertad para una constante unilateral izquierdo es $t_{\alpha, 5} = -2{,}015$.

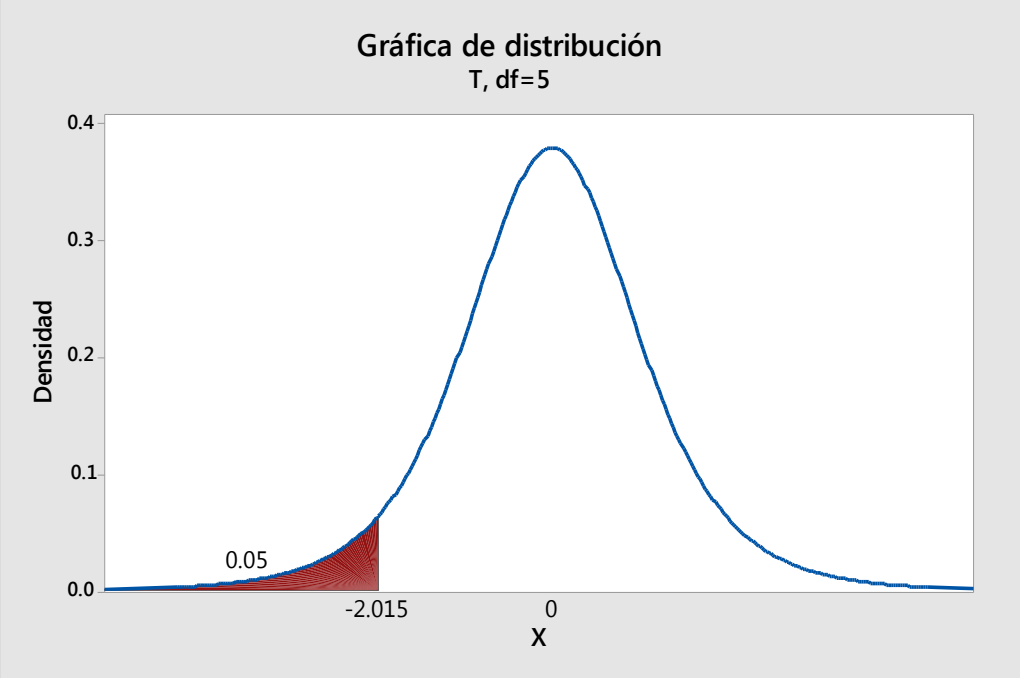


Figura 52. Gráfica de distribución para I3.

Conclusión:

Como el valor obtenido es t = -12,667 es menor que $t_{\alpha, 5} = -2{,}015$, se rechaza la hipótesis nula y queda demostrado que aplicando la técnica de minería de datos basado en árboles de decisión reduce la dificultad de la predicción del riesgo de morosidad de clientes en la empresa de seguros Oncosalud.

- **Contrastación para el tiempo para predecir**

  H_0: La aplicación de minería de datos basado en árboles de decisión no reduce el tiempo de predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

  H_1: La aplicación de minería de datos basado en árboles de decisión reduce el tiempo de predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

H_0: $\mu_1 \leq \mu_2$

H_1: $\mu_1 > \mu_2$

Donde:

- μ_1 = Media del tiempo para predecir el riesgo de morosidad con la Pre-Prueba.
- μ_2 = Media del tiempo para predecir el riesgo de morosidad con la Post-Prueba.

Tabla 38

Estadísticos descriptivos para el KPI4

	Pre-Prueba	**Post-Prueba**
Media (x)	689,45	58,70
Desviación estándar (S)	53,01	5,09
Observaciones (n)	382	382
Correlación	0,836	
Diferencia hipotética de las medias	630,74	
Gl	381	
Estadístico t	252,44	
p-valor (una cola)	0,000	

Región crítica

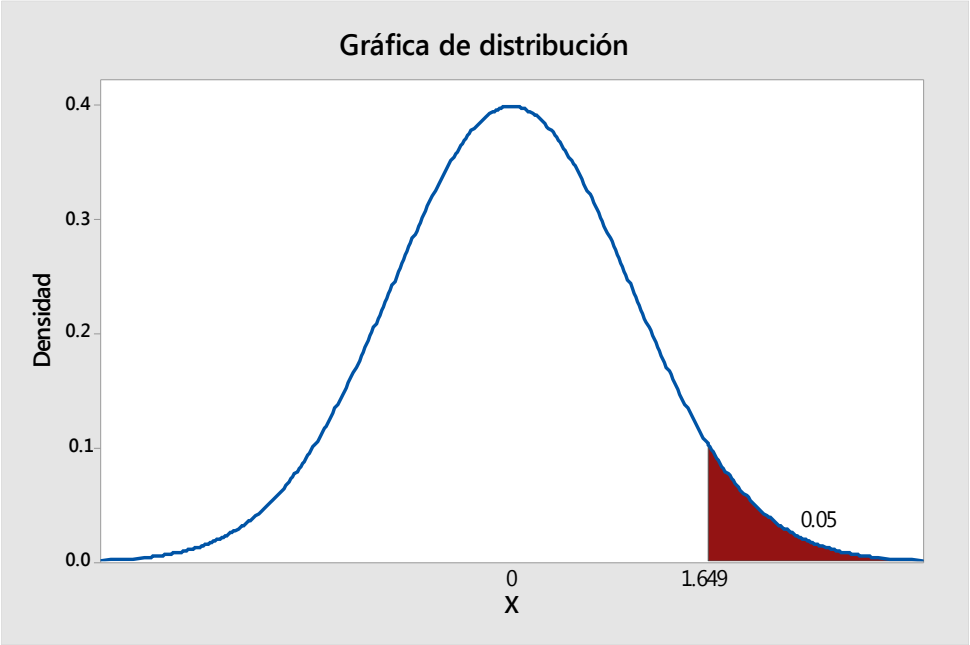


Figura 53. Gráfico de distribución para el I4.

Decisión estadística:

Puesto que el valor-p = 0.000 < ∝ = 0.05, se rechaza la hipótesis nula (H_0) y se acepta la hipótesis alterna (H_1).

Conclusión:

Se ha demostrado que el tiempo para predecir el riesgo de morosidad usando árboles de decisión es menor que el tiempo para predecir sin árboles de decisión con un 95% de probabilidad. Efectivamente la aplicación de minería de datos basado en árboles de decisión reduce significativamente el tiempo de predicción del riesgo de morosidad de la empresa de seguros Oncosalud.

CAPÍTULO V
CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- Se determinó que la aplicación de la técnica de minería de datos basado en árboles de decisión mejoró la precisión de predicción del riesgo de morosidad de los clientes en un 26% utilizando el algoritmo ID3 de árboles de decisión, en concordancia con Camborda (2014). Ver tabla 28.
- Se determinó que la aplicación de la técnica de minería de datos basado en árboles de decisión redujo la dificultad de la predicción del riesgo de morosidad en un 80% en los empleados del departamento de cobranza en comparación con el trabajo tradicional que se llevaba a diario en la empresa, en concordancia con Díaz (2016). Ver gráfico 47 y 48.
- Se determinó que la aplicación de la técnica de minería de datos basado en árboles de decisión redujo el tiempo de predicción del riesgo de morosidad de los clientes en un 91% en comparación al tiempo que se tomaban de la manera tradicional, en concordancia con Díaz (2016). Ver tabla 34.
- De manera general se concluye que la técnica de minería de datos basado en árboles de decisión facilitó la predicción del riesgo de morosidad de los clientes puesto que mejora la precisión de la predicción, reduce el nivel de dificultad y el tiempo para predecir tal como se mencionan en las conclusiones anteriores.

5.2 RECOMENDACIONES

- Se sugiere utilizar diferentes tipos de algoritmos de árboles de decisión como J48, Random forest, redes bayesianas entre otros para determinar quién tiene mayor grado de certeza para la predicción del riesgo de morosidad de clientes.
- Se sugiere que las reglas establecidas en el modelo desarrollado deben cambiarse anualmente ya que si siempre se utiliza las mismas reglas el grado de certeza en la predicción puede fallar o no ser tan preciso.
- Se sugiere que para este tipo de investigación no solo se debe tener conocimiento en un software si no también tener conocimiento en otros tipos de software que son utilizados en la minería de datos, por ejemplo, Rapidminer, R, Python entre otros, con el fin de ver cuál de ellos da mayor aporte a la investigación.
- Por último, se recomienda que para poder realizar la integración de JAVA con algún otro software de minería de datos se debe realizar la investigación necesaria para corroborar si se puede llevar a cabo esta integración del software de minería de datos con el entorno que se está desarrollando el sistema para no tener inconvenientes posteriormente, mayor detalle véase la figura 35.

REFERENCIAS BIBLIOGRÁFICAS

Artículos

Adeyemo, O., y Adeyeye, T. (2015). Comparative Study of ID3/C4.5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever. *African Journal of Computing & ICT, 8*(1), 103-112.

Altamiranda, L., Peña, A., Ospino, M., Volpe, I., Ortega, D., y Cantillo, E. (2013). Minería de datos como herramienta para el desarrollo de estrategias de mercadeo B2B en sectores productivos, afines a los colombianos: una revisión de casos. Sotavento M.B.A., (22), 126-136. Recuperado de https://revistas.uexternado.edu.co/index.php/sotavento/article/view/3709

Aranda, Y., y Sotolongo, A. (2013). Integración de los algoritmos de minería de datos 1R, PRISM E ID3 a Postgresql. *JISTEM – Journal of Information System and Technology Management, 10*(2), 389-406.

Arora, A., y Rajput, S. (2013). Designing Spam Model- Classification Analysis using Decision Trees. *International Journal of Computer Applications. 75*(10), 6-12.

Barrientos, R., Cruz, N., Acosta, H., Rabatte, I., Gogeascoechea, M., Pavón, P., y Blázquez, S. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista médica de la universidad veracruzana, 9*(2), 19-24.

Bhatt, H., Mehta, S., y D'mello, L. (2015). Use of ID3 Decision Tree Algorithm for Placement Prediction. *International Journal of Computer Science and Information Technologies, 6(5*), 4785-4789.

Chavarín, R. (2015). Morosidad en el pago de créditos y rentabilidad de la banca comercial en México. *Revista Mexicana de Economía y Finanzas, 10*(1), 73-85.

Contreras, E., Ferreira, F., y Valle, M. (2017). Diseño de un modelo predictivo de fuga de clientes utilizando árboles de decisión. *Ingeniería Industrial*, *16*(1), 7-23.

Fayyad, U., Piatetsky, G., y Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, *17*(3), 37-54.

Jaramillo, A., y Paz, H. (2015). Aplicación de técnicas de Minería de Datos para Determinar las Interacciones de los estudiantes en un entorno virtual de Aprendizaje. *Revista Tecnológica ESPOL – RTE, 28*(1), 64-90.

Kabari, L., y Bakpo, F. (2009). Credit Risk Evaluation System: An Artificial Neural Network Approach. *Nigerian Journal of Technology, 28*(1), 253-270.

Malviya, R., y Umrao, B. (2014). Comparison of NBTree and VFI Machine Learning Algorithms for Network Intrusion Detection using Feature Selection. *International Journal of Computer Applications, 108*(2), 35-38

Marulanda, C., López, M., y Mejía, M. (2017). Minería de datos en gestión del conocimiento de pymes de Colombia. Revista Virtual Universidad Católica del Norte, (50), 224-237. Recuperado de http://revistavirtual.ucn.edu.co/index.php/RevistaUCN/article/view/821/1339

Montequín, T., Álvarez, J., Mesa, J., y Gonzáles, A. (2003). Metodologías para la realización de proyectos de Data Mining. *AEIPRO, 1*(1), 257-265. Recuperado de https://www.aeipro.com/es/repository/func-startdown/2134/lang,es-es/

Paz, C., Ojeda, J., Badillo, E., Bonett, J., y Heredia, D. (2018). Reconocimiento de Dígitos Manuscritos por Medio de Técnicas de Minería de Datos. *Investigación y desarrollo en Tic, 8*(2), 46-50.

Solarte, G., y Soto, J. (2011). Árboles de decisión en el diagnóstico de enfermedades cardiovasculares. *Scientia Et Technica, 16*(49), 104-109.

Srivastava, S., y Joshi, N. (2014). Improving Classification Accuracy Using Ensemble Learning Technique (Using Different Decision Trees). *International Journal of Computer Science and Mobile Computing. 3*(5), 727-732.

Tello, M., Eslava, H., y Tobías, L. (2013). Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos. *Visión Electrónica, 7*(1), 13-26.

Vallejo, D., y Tenelada, G. (2012). Minería de datos aplicada en detección de intrusos. *Ing. USBMed, 3*(1), 50-61.

Wilford, I. (2006). Minería de datos: herramienta de apoyo en la selección de equipos de proyectos informáticos. *Ingeniería industrial, 27*(2-3), 7-10.

Tesis

Britos, P. (2008). *Procesos de explotación de información basados en sistemas inteligentes* (Tesis doctoral). Universidad nacional de la plata, Buenos Aires, Argentina.

Carpio, C. (2016). *Modelo de predicción de morosidad en el otorgamiento de crédito financiero aplicando metodología CRISP-DM* (Tesis de maestría). Universidad Andina Néstor Cáceres Velázquez, Juliaca, Perú.

Chero, K., y Paredes, M. (2016). *Estrategia crediticia para disminuir el índice de morosidad en el banco azteca, Chepen 2015* (Tesis de maestría). Universidad Señor de Sipán, Chiclayo, Perú.

Córdova, J. (2014). *Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes que pertenecen al colegio fisco-misional "San Francisco" de la ciudad de Ibarra* (Tesis de pregrado). Universidad Regional Autónoma de los Andes, Ibarra, Ecuador.

Díaz, A. (2016). *Técnicas de Minería de datos para predicción del diagnóstico de hipertensión arterial* (Tesis de pregrado). Universidad Señor de Sipán, Chiclayo, Perú.

Espino, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso* (Tesis de pregrado). Universidad Abierta de Cataluña, España.

Gopalakrishnan, A. (2016). *A multifaceted data mining approach to analyzing college students' persistence and graduation* (Tesis de maestría). San Francisco State University, San Francisco, California.

Martínez, M. (2013). *Gestión de riesgo en las entidades financieras: El riesgo de crédito y morosidad* (Tesis de pregrado). Universidad de Valladolid, España.

Ordoñez, K. (2013). *Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL - ECUADOR* (Tesis de pregrado). Universidad Técnica particular de Loja, Ecuador.

Puncernau, M. (2016). *Editor de árboles de decisión* (Tesis de pregrado). Universidad Politécnica de Cataluña, España.

Roque, I. (2016). *Análisis comparativo de técnicas de minería de datos para la predicción de ventas* (Tesis de pregrado). Universidad Señor de Sipán, Chiclayo, Perú.

SantaMaria, W. (2010). *Modelo de detección de fraude basado en el descubrimiento simbólico de reglas de clasificación extraídas de una red neuronal* (Tesis de pregrado). Universidad Nacional de Colombia, Bogotá, Colombia.

Sitios Web

Asbanc: morosidad bancaria subió a 2.96% en el primer mes del año. (19 de febrero de 2017). *SemanaEconomica.com*. Recuperado de http://semanaeconomica.com/article/mercados-y-finanzas/banca-y-finanzas/215406-asbanc-morosidad-bancaria-subio-a-2-96-en-el-primer-mes-del-ano/

Brachfield, P. (2014). *Las causas principales por las que existen morosos.* España: Brachfield Credit & Risk Consultants. Recuperado de http://perebrachfield.com/blog/morosos-y-pufistas/las-causas-principales-por-las-que-existen-morosos/

Brachfield, P. (2014). *Los seis grandes tipos de deudores.* España: Brachfield Credit & Risk Consultants. Recuperado de http://perebrachfield.com/blog/morosos-y-pufistas/los-seis-grandes-tipos-de-deudores/

DAS. (2015). *Claves para luchar contra la morosidad* [archivo PDF]. Recuperado de https://www.das.es/app/uploads/2017/01/Claves-para-luchar-contra-la-morosidad.pdf

Endara, C. (2006). *Gestión efectiva de cobranzas*. *5 claves de éxito. Bogotá,* Colombia: Gestiopolis. Recuperado de https://www.gestiopolis.com/gestion-efectiva-de-cobranzas-5-claves-de-exito/

Giraldo, W. (2010). *Determinantes de la morosidad de la cartera en el sistema financiero* [archivo PDF]. Recuperado de https://repository.icesi.edu.co/biblioteca_digital/bitstream/item/5394/1/Trabajo%20_Grado(WGY).pdf.

Gonzales, A. (2016). *Minería de datos* [archivo PDF]. Recuperado de https://ccc.inaoep.mx/~jagonzalez/AI/Sesion13_Data_Mining.pdf

Kobsa: Créditos crecen en 2.23% pero se reduce la capacidad de asumir una deuda. (16 de agosto de 2017). *Gestión*. Recuperado de https://gestion.pe/tu-dinero/kobsa-creditos-crecen-2-23-reduce-capacidad-asumir-deuda-141762

Larrieta, M. I. Á., y Gómez, A. M. S. (1998). Minería de datos: Concepto, características, estructura y aplicaciones. Recuperado de https://www.redib.org/recursos/Record/oai_articulo842354-miner%C3%ADa-datos-concepto-caracter%C3%ADsticas-estructura-aplicaciones/Bibliography#tabnav

León, E. (2007). *Módulo de minería de datos* [archivo PDF]. Recuperado de http://disi.unal.edu.co/~eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf

Lozano, M. (2011). El papel de las redes bayesianas en la toma de decisiones. Recuperado de https://docplayer.es/13694103-El-papel-de-las-redes-bayesianas-en-la-toma-de-decisiones-miller-rivera-lozano-miller-rivera-urosario-edu-co.html

Moine, J., Gordillo, S., y Haedo., A. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. Recuperado de http://hdl.handle.net/10915/18749

Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. KDnuggets. Recuperado de https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Vega, C., Rosano, G., López, J., Cendejas, J. y Ferreira, H. (2012). *Data Mining Aplicado a la Predicción y Tratamiento de Enfermedades* [archivo PDF]. Recuperado de http://www.iiis.org/cds2012/cd2012sci/cisci_2012/paperspdf/ca732ov.pdf

Westreicher, G. (26 de mayo de 2014). Deuda de las familias se extiende a 2.3 veces sus ingresos, según el BCR. *Gestión*. Recuperado de https://gestion.pe/economia/deuda-familias-extiende-2-3-veces-ingresos-bcr-61129

ANEXOS

Anexo I: Técnica del árbol

EFECTO:

Anexo II: Matriz de consistencia

Título: Aplicación de minería de datos basado en árboles de decisión para predecir el riesgo de morosidad de los clientes en la empresa de seguros Oncosalud S.A.C. 2018.

Problema Principal	Objetivos	Hipótesis	Variables	
General ¿En qué medida la aplicación de minería de datos basado en árboles de decisión facilitará la predicción del riesgo de morosidad de los clientes en la empresa de seguros Oncosalud? **Específicos:** *¿En qué medida la aplicación de minería de datos basado en árboles de decisión mejorará la precisión de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud? *¿En qué medida la aplicación de minería de datos basado en árboles de decisión reducirá la dificultad de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud? *¿En qué medida la aplicación de minería de datos basado en árboles de decisión reducirá el tiempo de predicción del riesgo de morosidad de la empresa de seguros Oncosalud?	**General** Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión facilita la predicción del riesgo de morosidad de los clientes en la empresa de seguros Oncosalud SAC 2018. **Específicos:** *Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión mejora la precisión de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud. *Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión reduce la dificultad de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud. *Determinar en qué medida la aplicación de minería de datos basado en árboles de decisión reduce el tiempo de predicción del riesgo de morosidad de la empresa de seguros Oncosalud.	**General** La aplicación de minería de datos basado en árboles de decisión facilita significativamente la predicción del riesgo de morosidad de los clientes en la empresa de seguros Oncosalud SAC 2018. **Específicos:** *La aplicación de minería de datos basado en árboles de decisión mejora significativamente la precisión de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud. *La aplicación de minería de datos basado en árboles de decisión reduce significativamente la dificultad de la predicción del riesgo de morosidad de la empresa de seguros Oncosalud. *La aplicación de minería de datos basado en árboles de decisión reduce significativamente el tiempo de predicción del riesgo de morosidad de la empresa de seguros Oncosalud.	**Independiente:** X: Árboles de decisión **Dependiente:** Y: Predicción del riesgo de morosidad	**Tipo de investigación:** Aplicada **Nivel de investigación:** Explicativa **Diseño de investigación:** Pre-Experimental **Población:** 85 000 clientes **Muestra:** 382 clientes

Anexo III: Matriz de Operacionalización de Variables

VARIABLE	DIMENSIÓN	CONCEPTUALIZACIÓN	INDICADORES	ÍNDICE	TÉCNICA	INSTRUMENTO
Predicción del riesgo de morosidad Conceptualización: Es la posibilidad de impago de una operación	Precisión	Es el grado de coincidencia existente entre los resultados independientes de una medición.	Precisión de predicción	[60%,...,95%]	Revisión de documentos Software	Reporte
			Error de predicción	[1%,...,40%]	Revisión de documentos Software	Reporte
	Dificultad	Según Pérez y Merino (2008) definen problema que surge cuando una persona intenta lograr algo, por lo tanto son inconvenientes o barreras que hay que superar para conseguir un determinado objetivo.	Nivel de dificultad	[Muy difícil, Difícil, Normal, Fácil, Muy fácil]	Encuesta	Cuestionario
	Tiempo	El tiempo es la magnitud física que permite secuenciar hechos y determinar momentos y cuya unidad de medida es el segundo.	Tiempo para predecir	[1 - 780]	Observación directa con cronometro	Ficha de registro

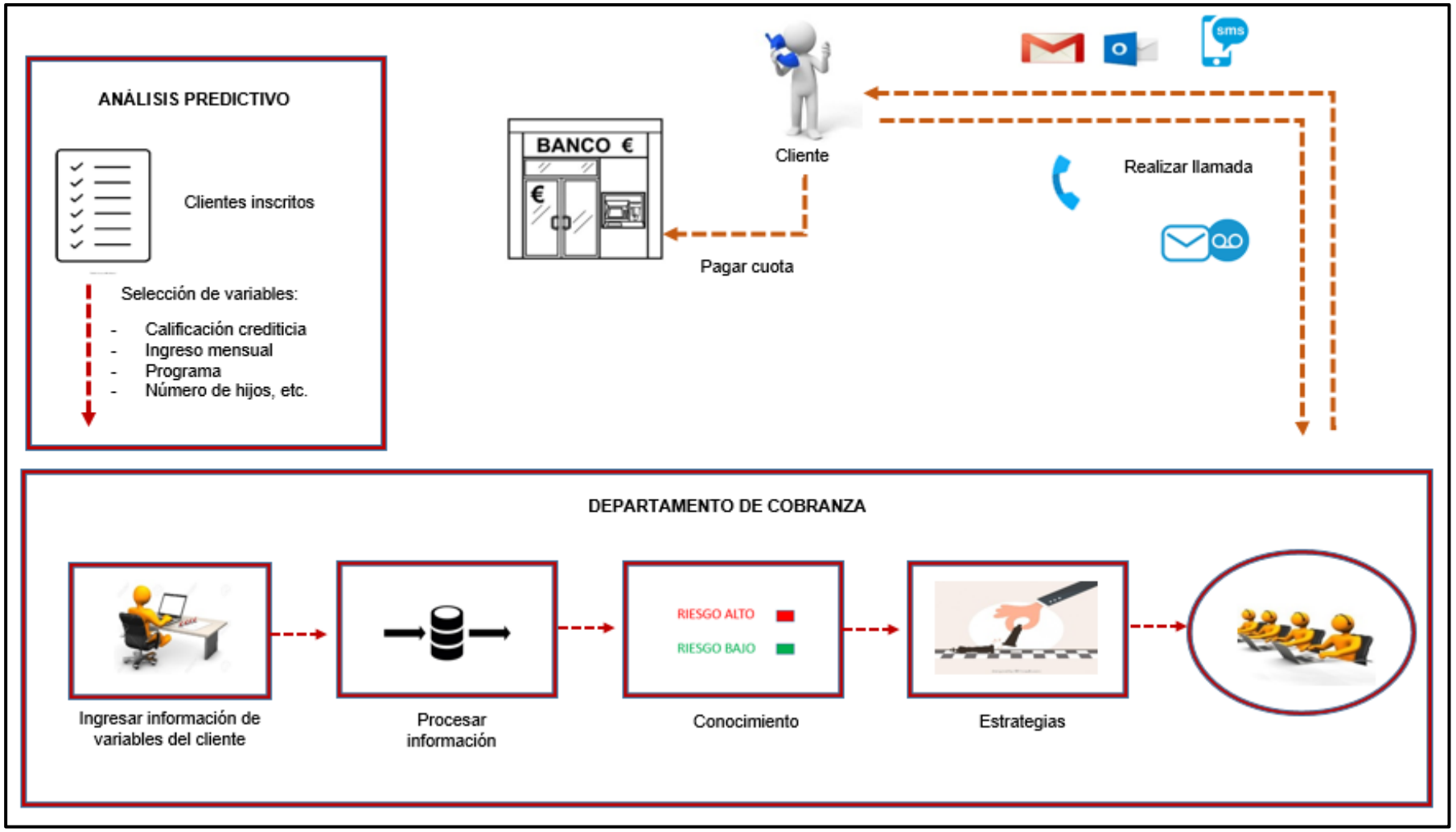
Anexo IV: Matriz de revisión de antecedentes

Matriz de revisión				
Título	**Autor**	**Universidad**	**Año**	**Fecha de revisión**
Comparación interactiva de modelos de minería de datos utilizando técnicas de visualización.	Padua, Luciana María	Universidad de buenos aires – Argentina	2014	23/10/17
Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes que pertenecen al colegio fiscomisional “san francisco” de la ciudad de Ibarra.	Córdova Galarza Janeth Carolina	Universidad Regional Autónoma de los Andes “Uniandes” – Ecuador	2014	23/10/17
Modelo de RNA para predecir la morosidad de microcrédito en la Banca Estatal Peruana.	Luis Esteban Palacios Quichiz	Universidad Nacional Mayor de San Marcos	2012	23/10/17
Técnicas de Minería de datos para predicción del diagnóstico de hipertensión arterial.	Díaz Avendaño, Ángel Arnulfo	Universidad Señor de Sipán – Perú	2016	24/10/17
Modelo de predicción de la morosidad en el otorgamiento de crédito financiero aplicando metodología CRISP-DM.	Julio Cesar Carpio Ticona	Universidad Andina Néstor Cáceres Velázquez – Juliaca	2016	24/10/17
Implementación de un sistema de información basado en un enfoque de procesos, para la mejora de la operatividad del área de créditos de la Microfinanciera Crecer.	Bendezu Tenorio Natalu Delia	Universidad Nacional del centro del Perú – Huancayo	2014	24/10/17
Estrategias crediticias para disminuir el índice de morosidad en el banco Azteca, Chepen 2015	Chero Vásquez, Keysi y Paredes Abanto, María	Universidad Señor de Sipán - Perú	2016	24/10/17

Anexo V: Matriz de revisión de artículos

Matriz de Revisión de Artículos				
Nombre artículo	**Año**	**Autor(es)**	**País/universidad**	**Tema del articulo**
Análisis y evaluación de los riesgo financieros en la cooperativa de ahorro y crédito Kullki Wasi Ltda. De la ciudad de Ambato, periodo 2011-2014 y rediseño de un plan estratégico para optimizar la gestión de los riesgos.	2015	Jessica Chiluisa Luis Tenelesma Marco Veloz	Universidad de las fuerzas armadas ESPE – Ecuador	Mide los riesgos financieros que permitirá evaluar en qué situación se encuentran las cooperativas de ahorro y crédito del sector popular y solidario, realiza un rediseño a la planificación estratégica con la optimización de estrategias financieras.
Reconocimiento de Dígitos Manuscritos por Medio de Técnicas de Minería de Datos.	2018	Paz, C., Ojeda, J., Badillo, E., Bonett, J., y Heredia, D	Universidad Simón Bolívar - Colombia	Aplicación de técnicas de minería de datos, árboles de decisión, para el reconocimiento de dígitos manuscritos.
Aplicación de técnicas de Minería de Datos para Determinar las Interacciones de los estudiantes en un entorno virtual de Aprendizaje.	2015	Angélica Jaramillo Henry Paz	Universidad Nacional de Loja - Ecuador	Determina las interacciones de los estudiantes para seleccionar los atributos necesarios para generar un modelo de minería de datos.
Árboles de decisión como herramienta en el diagnóstico médico.	2009	Barrientos, R., Cruz, N., Acosta, H., Rabatte, I., Gogeascoechea, M., Pavón, P., y Blázquez, S.	Universidad Veracruzana - México	Evalúa el desempeño de tres algoritmos para la construcción de árboles de decisiones con respecto al diagnóstico de cáncer de seno.
Minería de datos aplicada en detección de intrusos.	2012	Vallejo, D., Tenelada, G.	Universidad de Colombia	Muestra el aporte a la seguridad de la información de la minería de datos en el contexto de la detección de intrusos utilizando la metodología CRISP-DM

Anexo VI: Modelo de la solución



Anexo VII: Contención de clientes de la empresa Oncosalud

	2018-01		2018-02		2018-03		2018-04	
Rango de mora inicial	**Porcentaje de contención S/.**	**Cantidad de grupo familiar**	**Porcentaje de contención S/.**	**Cantidad de grupo familiar**	**Porcentaje de contención S/.**	**Cantidad de grupo familiar**	**Porcentaje de contención S/.**	**Cantidad de grupo familiar**
A.[Al día]	94.3%	269,843	93.0%	265,919	95.1%	275,737	94.3%	271,941
B.[1-30 días]	45.5%	9,335	41.6%	7,974	46.3%	12,159	45.4%	9,022
C.[31-60 días]	24.6%	2,024	22.2%	1,882	22.1%	1,517	26.8%	2,662
D.[61-90 días]	56.4%	799	45.4%	596	45.4%	644	57.0%	682
E.[91-120 días]	64.3%	340	51.7%	217	41.7%	173	54.0%	239
F.[121-150 días]	82.6%	49	43.5%	7	46.0%	14	57.5%	6
G.[+ 180 días]	76.9%	2	1.7%	17	71.2%	18	28.6%	14

Nota: Periodo 2018-01 al 2018-04

Anexo VIII: Resultados del algoritmo ID3 de árboles de decisión

Clase	Precisión	Recall	ROC Área
RIESGO ALTO	0.875	0.875	0.891
RIESGO BAJO	0.905	0.923	0.89

Anexo IX: Algoritmo de árboles de decisión

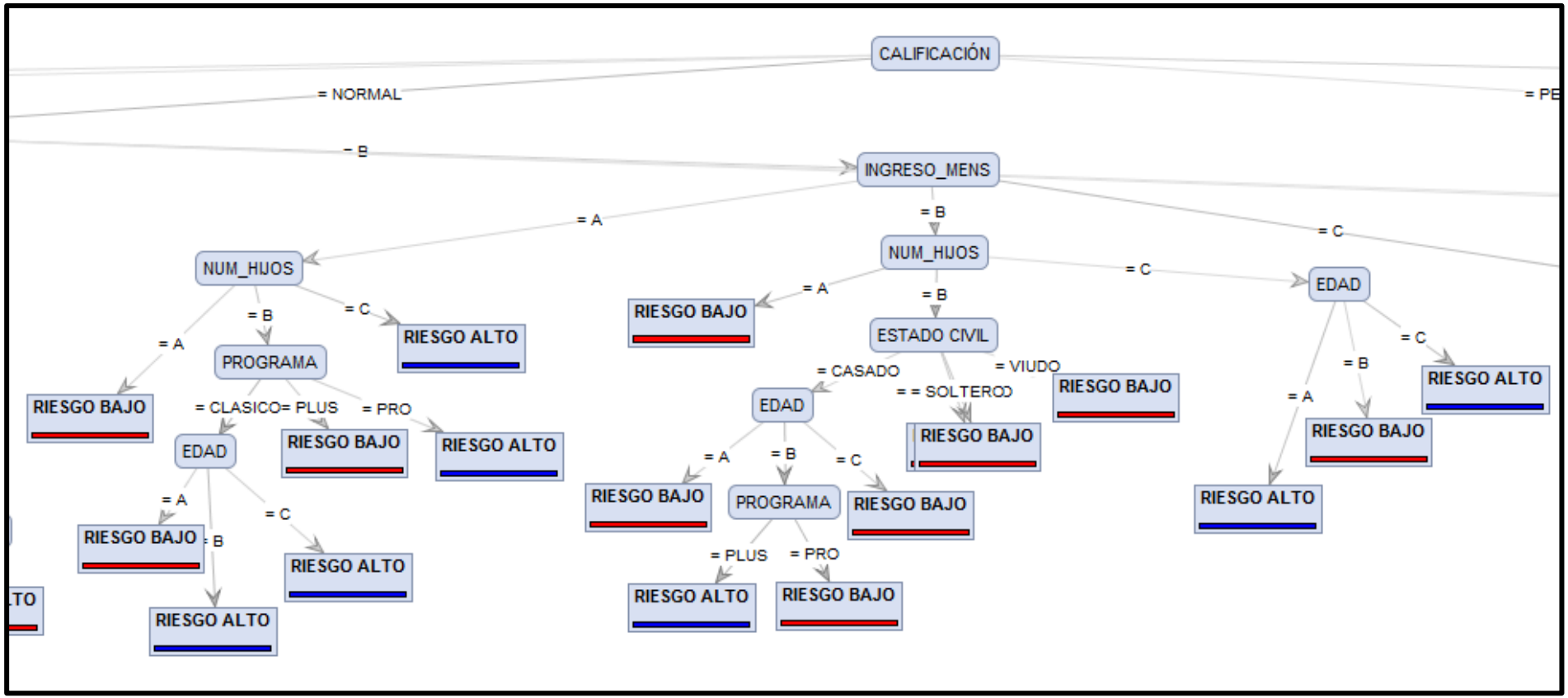
```
Id3


CALIFICACIÓN = PROBLEMAS POTENCIALES
|  DEUDA_ONCO = C: RIESGO ALTO
|  DEUDA_ONCO = A
|  |  INGRESO_MENS = B
|  |  |  NUM_HIJOS = C: RIESGO ALTO
|  |  |  NUM_HIJOS = A: RIESGO BAJO
|  |  |  NUM_HIJOS = B
|  |  |  |  ESTADO CIVIL = SOLTERO
|  |  |  |  |  PROGRAMA = PLUS: RIESGO ALTO
|  |  |  |  |  PROGRAMA = CLASICO: RIESGO BAJO
|  |  |  |  |  PROGRAMA = PRO: RIESGO BAJO
|  |  |  |  ESTADO CIVIL = DIVORCIADO: RIESGO BAJO
|  |  |  |  ESTADO CIVIL = VIUDO: RIESGO BAJO
|  |  |  |  ESTADO CIVIL = CASADO: RIESGO BAJO
|  |  INGRESO_MENS = C
|  |  |  ESTADO CIVIL = SOLTERO: RIESGO BAJO
|  |  |  ESTADO CIVIL = DIVORCIADO
|  |  |  |  EDAD = A: RIESGO BAJO
|  |  |  |  EDAD = C
|  |  |  |  |  PROGRAMA = PLUS: RIESGO ALTO
|  |  |  |  |  PROGRAMA = CLASICO: RIESGO BAJO
|  |  |  |  |  PROGRAMA = PRO: null
|  |  |  |  EDAD = B: RIESGO BAJO
|  |  |  ESTADO CIVIL = VIUDO: RIESGO BAJO
|  |  |  ESTADO CIVIL = CASADO: RIESGO BAJO
|  |  INGRESO_MENS = A
|  |  |  NUM_HIJOS = C
|  |  |  |  ESTADO CIVIL = SOLTERO: RIESGO ALTO
|  |  |  |  ESTADO CIVIL = DIVORCIADO: RIESGO ALTO
```

```
|   |   |   |  ESTADO CIVIL = VIUDO
|   |   |   |   |  PROGRAMA = PLUS
|   |   |   |   |   |  EDAD = A: null
|   |   |   |   |   |  EDAD = C: RIESGO ALTO
|   |   |   |   |   |  EDAD = B: RIESGO BAJO
|   |   |   |   |  PROGRAMA = CLASICO: RIESGO ALTO
|   |   |   |   |  PROGRAMA = PRO: RIESGO BAJO
|   |   |   |  ESTADO CIVIL = CASADO: RIESGO ALTO
|   |   |  NUM_HIJOS = A
|   |   |   |  ESTADO CIVIL = SOLTERO: RIESGO BAJO
|   |   |   |  ESTADO CIVIL = DIVORCIADO
|   |   |   |   |  EDAD = A: RIESGO BAJO
|   |   |   |   |  EDAD = C: null
|   |   |   |   |  EDAD = B: RIESGO ALTO
|   |   |   |  ESTADO CIVIL = VIUDO: null
|   |   |   |  ESTADO CIVIL = CASADO: RIESGO BAJO
|   |   |  NUM_HIJOS = B
|   |   |   |  PROGRAMA = PLUS: RIESGO ALTO
|   |   |   |  PROGRAMA = CLASICO: RIESGO BAJO
|   |   |   |  PROGRAMA = PRO
|   |   |   |   |  EDAD = A: RIESGO ALTO
|   |   |   |   |  EDAD = C
|   |   |   |   |   |  ESTADO CIVIL = SOLTERO: RIESGO BAJO
|   |   |   |   |   |  ESTADO CIVIL = DIVORCIADO: RIESGO ALTO
|   |   |   |   |   |  ESTADO CIVIL = VIUDO: null
|   |   |   |   |   |  ESTADO CIVIL = CASADO: null
|   |   |   |   |  EDAD = B: RIESGO BAJO
|  DEUDA_ONCO = B
|   |  INGRESO_MENS = B
|   |   |  ESTADO CIVIL = SOLTERO
|   |   |   |  EDAD = A
|   |   |   |   |  NUM_HIJOS = C: null
|   |   |   |   |  NUM_HIJOS = A: RIESGO BAJO
|   |   |   |   |  NUM_HIJOS = B: RIESGO ALTO
|   |   |   |  EDAD = C: RIESGO ALTO
|   |   |   |  EDAD = B: RIESGO ALTO
|   |   |  ESTADO CIVIL = DIVORCIADO
|   |   |   |  NUM_HIJOS = C: RIESGO ALTO
|   |   |   |  NUM_HIJOS = A: RIESGO BAJO
|   |   |   |  NUM_HIJOS = B: RIESGO BAJO
|   |   |  ESTADO CIVIL = VIUDO
|   |   |   |  PROGRAMA = PLUS: RIESGO BAJO
|   |   |   |  PROGRAMA = CLASICO: RIESGO BAJO
|   |   |   |  PROGRAMA = PRO: RIESGO ALTO
|   |   |  ESTADO CIVIL = CASADO
|   |   |   |  PROGRAMA = PLUS: null
|   |   |   |  PROGRAMA = CLASICO: RIESGO ALTO
```

```
|   |   |   |   |  PROGRAMA = PRO: RIESGO BAJO
|   |   |   |  ESTADO CIVIL = CASADO: RIESGO ALTO
|   |   |  NUM_HIJOS = A
|   |   |   |  ESTADO CIVIL = SOLTERO: RIESGO BAJO
|   |   |   |  ESTADO CIVIL = DIVORCIADO
|   |   |   |   |  EDAD = A: RIESGO BAJO
|   |   |   |   |  EDAD = C: null
|   |   |   |   |  EDAD = B: RIESGO ALTO
|   |   |   |  ESTADO CIVIL = VIUDO: null
|   |   |   |  ESTADO CIVIL = CASADO: RIESGO BAJO
|   |   |  NUM_HIJOS = B
|   |   |   |  PROGRAMA = PLUS: RIESGO ALTO
|   |   |   |  PROGRAMA = CLASICO: RIESGO BAJO
|   |   |   |  PROGRAMA = PRO
|   |   |  INGRESO_MENS = B
|   |   |   |  ESTADO CIVIL = SOLTERO: RIESGO ALTO
|   |   |   |  ESTADO CIVIL = DIVORCIADO: RIESGO BAJO
|   |   |   |  ESTADO CIVIL = VIUDO: null
|   |   |   |  ESTADO CIVIL = CASADO
|   |   |   |   |  EDAD = A: RIESGO ALTO
|   |   |   |   |  EDAD = C: RIESGO BAJO
|   |   |   |   |  EDAD = B: RIESGO ALTO
|   |   |  INGRESO_MENS = C: RIESGO BAJO
|   |   |  INGRESO_MENS = A
|   |   |   |  PROGRAMA = PLUS: RIESGO ALTO
|   |   |   |  PROGRAMA = CLASICO: RIESGO BAJO
|   |   |   |  PROGRAMA = PRO: RIESGO ALTO
|   |  NUM_HIJOS = B
|   |   |  INGRESO_MENS = B
|   |   |   |  EDAD = A: RIESGO ALTO
|   |   |   |  EDAD = C: RIESGO BAJO
|   |   |   |  EDAD = B: null
|   |   |  INGRESO_MENS = C: RIESGO BAJO
|   |   |  INGRESO_MENS = A: RIESGO ALTO
|  DEUDA_ONCO = B
|   |  INGRESO_MENS = B: RIESGO ALTO
|   |  INGRESO_MENS = C
|   |   |  NUM_HIJOS = C
|   |   |   |  PROGRAMA = PLUS: RIESGO ALTO
|   |   |   |  PROGRAMA = CLASICO: RIESGO ALTO
|   |   |   |  PROGRAMA = PRO: RIESGO ALTO
|   |   |  NUM_HIJOS = A: RIESGO BAJO
|   |   |  NUM_HIJOS = B: RIESGO ALTO
|   |  INGRESO_MENS = A: RIESGO ALTO
|  DEUDA_ONCO = D: RIESGO ALTO
CALIFICACIÓN = PERDIDAS: RIESGO ALTO
```

Anexo X: Arboles de decisión ID3 en RapidMiner



Anexo XI: Resultados del Indicador Tiempo para predecir (KPI4)

	KPI4: Tiempo para predecir												
N°	**Pre-Prueba**	**Post-Prueba**	**N°**	**Pre-Prueba**	**Post-Prueba**	**N°**	**Pre-Prueba**	**Post-Prueba**	**N°**	**Pre-Prueba**	**Post-Prueba**	**N°**	**Pre-Prueba**	**Post-Prueba**
1	740	66	**27**	688	55	**53**	675	52	**79**	641	58	**105**	604	50
2	639	53	**28**	643	54	**54**	619	50	**80**	626	53	**106**	663	53
3	628	53	**29**	688	53	**55**	757	70	**81**	720	63	**107**	707	61
4	768	60	**30**	705	53	**56**	724	69	**82**	629	60	**108**	719	62
5	749	57	**31**	649	51	**57**	688	53	**83**	770	67	**109**	722	62
6	627	53	**32**	600	50	**58**	685	53	**84**	678	59	**110**	737	63
7	647	56	**33**	673	53	**59**	697	54	**85**	618	55	**111**	772	65
8	654	58	**34**	601	50	**60**	629	53	**86**	620	55	**112**	708	60
9	672	60	**35**	705	53	**61**	664	55	**87**	705	61	**113**	650	53
10	713	65	**36**	768	68	**62**	602	50	**88**	728	63	**114**	716	61
11	673	60	**37**	744	60	**63**	741	67	**89**	780	65	**115**	636	58
12	740	55	**38**	625	51	**64**	774	69	**90**	619	51	**116**	651	58
13	638	53	**39**	675	52	**65**	725	64	**91**	702	61	**117**	637	56
14	739	55	**40**	667	51	**66**	688	53	**92**	665	53	**118**	627	56
15	618	52	**41**	603	50	**67**	624	51	**93**	779	62	**119**	771	65
16	778	67	**42**	727	59	**68**	778	69	**94**	705	58	**120**	768	64
17	656	58	**43**	663	57	**69**	660	52	**95**	623	52	**121**	666	53
18	667	57	**44**	616	51	**70**	641	53	**96**	715	61	**122**	768	65
19	757	56	**45**	600	50	**71**	687	54	**97**	738	61	**123**	744	66
20	730	55	**46**	754	63	**72**	709	61	**98**	739	61	**124**	708	61
21	717	53	**47**	738	58	**73**	717	62	**99**	623	52	**125**	729	63
22	609	50	**48**	665	57	**74**	716	62	**100**	692	55	**126**	739	63
23	650	51	**49**	650	56	**75**	652	58	**101**	627	53	**127**	610	50
24	751	67	**50**	779	69	**76**	765	67	**102**	679	55	**128**	687	55
25	673	50	**51**	748	62	**77**	735	65	**103**	631	54	**129**	684	55
26	608	50	**52**	603	50	**78**	656	54	**104**	760	60	**130**	776	65

131	604	50	**160**	654	55	**189**	633	53	**218**	733	63	**247**	780	70
132	650	53	**161**	774	69	**190**	733	63	**219**	645	53	**248**	626	51
133	766	63	**162**	691	59	**191**	694	59	**220**	701	60	**249**	706	60
134	674	54	**163**	711	65	**192**	631	52	**221**	615	51	**250**	632	52
135	734	64	**164**	675	58	**193**	737	62	**222**	612	51	**251**	758	66
136	600	50	**165**	629	56	**194**	655	55	**223**	709	60	**252**	702	60
137	707	60	**166**	714	65	**195**	710	60	**224**	716	61	**253**	600	50
138	608	50	**167**	712	65	**196**	723	61	**225**	668	57	**254**	610	50
139	765	62	**168**	616	51	**197**	665	55	**226**	756	64	**255**	701	60
140	714	60	**169**	778	69	**198**	688	58	**227**	636	52	**256**	718	61
141	705	60	**170**	614	51	**199**	678	57	**228**	627	52	**257**	646	52
142	735	62	**171**	720	62	**200**	655	55	**229**	649	52	**258**	765	66
143	726	62	**172**	681	58	**201**	703	60	**230**	752	64	**259**	630	61
144	726	62	**173**	616	51	**202**	614	50	**231**	725	62	**260**	672	67
145	726	62	**174**	702	60	**203**	643	53	**232**	759	65	**261**	682	67
146	674	57	**175**	779	66	**204**	743	63	**233**	766	66	**262**	706	62
147	656	56	**176**	624	52	**205**	697	58	**234**	689	56	**263**	770	63
148	775	65	**177**	629	52	**206**	642	55	**235**	751	65	**264**	740	62
149	766	64	**178**	700	60	**207**	705	60	**236**	670	57	**265**	627	52
150	747	63	**179**	755	65	**208**	755	63	**237**	698	58	**266**	670	56
151	622	50	**180**	759	65	**209**	692	58	**238**	678	57	**267**	761	64
152	693	59	**181**	657	55	**210**	618	50	**239**	649	53	**268**	668	54
153	609	50	**182**	680	56	**211**	734	62	**240**	631	52	**269**	681	56
154	715	60	**183**	664	54	**212**	640	53	**241**	626	51	**270**	619	51
155	619	51	**184**	778	67	**213**	776	66	**242**	749	65	**271**	735	62
156	754	64	**185**	632	53	**214**	686	56	**243**	658	54	**272**	699	59
157	623	52	**186**	718	61	**215**	719	60	**244**	750	64	**273**	770	69
158	677	55	**187**	775	67	**216**	671	56	**245**	736	62	**274**	624	51
159	659	54	**188**	774	67	**217**	772	65	**246**	730	62	**275**	687	57

276	701	50	**298**	764	**66**	**320**	616	51	**342**	603	50	**364**	694	59
277	702	50	**299**	749	**62**	**321**	748	62	**343**	758	65	**365**	672	56
278	650	52	**300**	649	**52**	**322**	689	56	**344**	637	60	**366**	771	70
279	777	66	**301**	664	**55**	**323**	634	63	**345**	732	65	**367**	616	57
280	602	60	**302**	768	**68**	**324**	687	57	**346**	755	63	**368**	639	57
281	721	64	**303**	631	**52**	**325**	606	51	**347**	657	53	**369**	731	62
282	631	55	**304**	727	**53**	**326**	622	67	**348**	610	50	**370**	752	65
283	776	59	**305**	614	**55**	**327**	721	61	**349**	731	62	**371**	678	54
284	637	55	**306**	629	**55**	**328**	738	66	**350**	691	58	**372**	609	51
285	683	53	**307**	679	**56**	**329**	656	67	**351**	760	65	**373**	742	65
286	709	69	**308**	638	**54**	**330**	618	52	**352**	634	54	**374**	632	53
287	671	63	**309**	747	**69**	**331**	758	63	**353**	635	55	**375**	668	55
288	647	59	**310**	762	**65**	**332**	623	51	**354**	689	68	**376**	646	57
289	695	56	**311**	683	**62**	**333**	646	52	**355**	656	69	**377**	749	58
290	624	62	**312**	742	**63**	**334**	780	65	**356**	678	53	**378**	673	65
291	611	66	**313**	713	**53**	**335**	779	64	**357**	746	70	**379**	623	52
292	707	63	**314**	676	**70**	**336**	655	55	**358**	752	70	**380**	662	62
293	765	64	**315**	698	**56**	**337**	730	61	**359**	675	56	**381**	600	51
294	751	64	**316**	655	**52**	**338**	619	57	**360**	727	62	**382**	726	61
295	695	59	**317**	714	**62**	**339**	620	54	**361**	745	61			
296	673	58	**318**	726	**65**	**340**	746	58	**362**	716	66			
297	759	65	**319**	750	**64**	**341**	633	55	**363**	656	54			

Anexo XII: Resultados obtenidos al aplicar el cuestionario a los empleados del departamento de cobranza.

CRITERIO	ESCALA
Muy difícil	1
Difícil	2
Normal	3
Fácil	4
Muy fácil	5

- **Resultado Pre-prueba**

ITEM	P1	P2	P3	P4	P5	TOTAL
TRABAJADOR1	2	2	2	2	2	10
TRABAJADOR2	1	1	1	1	2	6
TRABAJADOR3	2	2	2	1	2	9
TRABAJADOR4	2	2	1	2	2	9
TRABAJADOR5	1	1	1	1	1	5
TRABAJADOR6	2	1	2	2	2	9

- **Resultados Post-prueba**

ITEM	P1	P2	P3	P4	P5	TOTAL
TRABAJADOR1	4	4	4	5	4	21
TRABAJADOR2	5	4	4	4	4	21
TRABAJADOR3	5	5	4	5	5	24
TRABAJADOR4	5	4	5	4	5	23
TRABAJADOR5	5	5	5	5	5	25
TRABAJADOR6	5	5	4	5	5	24

APÉNDICE

APÉNDICE I: Cuestionario

CUESTIONARIO

El presente cuestionario tiene como objetivo identificar el nivel de dificultad de los empleados del departamento de cobranza respecto a la predicción del riesgo de morosidad de los clientes de la empresa Oncosalud.

Este cuestionario consta de 5 preguntas, lea atentamente cada una de ellas y responda.

Marca con un **aspa** o un **check** la opción más apropiada que considere.

Nombre:……

1: Muy difícil 2: Difícil 3: Normal 4: Fácil 5: Muy fácil

		1	2	3	4	5
1	¿Cómo considera usted la recolección de la información necesaria para determinar si un cliente podría ser moroso?					
2	¿Cómo considera usted al momento de avaluar la información del cliente?					
3	¿Cómo considera usted al momento de identificar si un cliente podría ser moroso?					
4	¿Cómo considera usted al momento de tomar decisiones frente a la información del cliente que ha sido evaluado?					
5	¿Cómo considera usted el manejo de las actividades diarias en el área de cobranza?					

APÉNDICE II: Ficha de registro

Aspecto a observar:

I4: Tiempo para predecir

Asegurados	Pre-prueba	Post-prueba
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
: .		
381		
382		

APENDICE III: Carta de aceptación para la realización de proyecto de investigación

CARTA DE ACEPTACIÓN PARA REALIZACIÓN DE PROYECTO DE INVESTIGACIÓN EN ONCOSALUD SAC

Lima 05 de diciembre de 2017

Sr.
José Luis Herrera Salazar
Director de Carrera Profesional de Ingeniería de Sistemas
Facultad de Ingeniería y Arquitectura
Universidad Autónoma del Perú
Presente. -

De nuestra consideración

Es grato dirigirme a ustedes en representación de ONCOSALUD S.A.C. para hacer de su conocimiento que la señorita Garcia Torres Maria Emily y el señor Espino Quiñones Leonardo, estudiantes de la carrera profesional de ingeniería de sistemas de vuestra institución universitaria Autónoma del Perú que usted representa, ha sido admitido para realizar su proyecto de tesis "TECNICAS DE MINERIA DE DATOS PARA PREDECIR EL RIESGO DE MOROSIDAD DE LOS CLIENTES EN LA EMPRESA DE SEGUROS ONCOSALUD SAC 2018" en la Gerencia de Gestión de Clientes de nuestra organización, teniendo como fecha de inicio el 01 de Diciembre del 2017.

Sin otro particular, quedo de usted

Atentamente

Aldo Ramirez Salas
Jefe de Recupero de Clientes

APÉNDICE IV: VALIDACIÓN DE JUICIO DE EXPERTOS

CERTIFICADO DE VALIDEZ DE CONTENIDO DE INSTRUMENTOS A TRAVÉS DE JUICIO DE EXPERTO

Título de la investigación	Aplicación de minería de datos basado en árboles de decisión para predecir el riesgo de morosidad de los clientes en la empresa de seguros Oncosalud S.A.C. 2018
Nombre(s) del(os) instrumento(s)	Cuestionario, Ficha de registro
Autor(es) del instrumento	Leonardo Espino Quiñones Maria Emily Garcia Torres

Nº	DIMENSIONES / Indicadores	Pertinencia[1]		Relevancia[2]		Claridad[3]		Sugerencias
DIMENSIÓN 1: Precisión		Si	No	Si	No	Si	No	
1	Precisión de predicción	X		X		X		
2	Error de predicción	X		X		X		
DIMENSIÓN 2: Dificultad		Si	No	Si	No	Si	No	
3	¿Cómo considera usted la recolección de la información necesaria para determinar si un cliente podría ser moroso?	X		X		X		
4	¿Cómo considera usted la evaluación de la información del cliente?	X		X		X		
5	¿Cómo considera usted la identificación de un cliente moroso?	X		X		X		
6	¿Cómo considera usted la toma de decisiones frente a la información del cliente que ha sido evaluado?	X		X		X		
7	¿Cómo considera usted las actividades diarias que se realiza en el departamento de cobranza?	X		X		X		
DIMENSIÓN 3: Tiempo		Si	No	Si	No	Si	No	
8	Tiempo para predecir	X		X		X		

Observaciones (precisar si hay suficiencia): .....

Opinión de aplicabilidad: **Aplicable [X]** **Aplicable después de corregir [ ]** **No aplicable [ ]**

Apellidos y nombres del juez validador. Dr/ Mg: Guevara Baca Victor **DNI:** 45422813

Especialidad del validador: Ing. Estadística

Cel: 986397261

03 de 12 del 2018

[1]**Pertinencia:** El ítem corresponde al concepto teórico formulado.
[2]**Relevancia:** El ítem es apropiado para representar al componente o dimensión específica del constructo
[3]**Claridad:** Se entiende sin dificultad alguna el enunciado del ítem, es conciso, exacto y directo

Nota: Suficiencia, se dice suficiencia cuando los ítems planteados son suficientes para medir la dimensión

Firma del Experto Informante.

CERTIFICADO DE VALIDEZ DE CONTENIDO DE INSTRUMENTOS A TRAVÉS DE JUICIO DE EXPERTO

Título de la investigación	**Aplicación de minería de datos basado en árboles de decisión para predecir el riesgo de morosidad de los clientes en la empresa de seguros Oncosalud S.A.C. 2018**
Nombre(s) del(os) instrumento(s)	**Cuestionario, Ficha de registro**
Autor(es) del instrumento	**Leonardo Espino Quiñones** **Maria Emily Garcia Torres**

Nº	DIMENSIONES / Indicadores	Pertinencia[1]		Relevancia[2]		Claridad[3]		Sugerencias
DIMENSIÓN 1: Precisión		Si	No	Si	No	Si	No	
1	Precisión de predicción	✓		✓		✓		
2	Error de predicción	✓		✓		✓		
DIMENSIÓN 2: Dificultad		Si	No	Si	No	Si	No	
3	¿Cómo considera usted la recolección de la información necesaria para determinar si un cliente podría ser moroso?	✓		✓		✓		
4	¿Cómo considera usted la evaluación de la información del cliente?	✓		✓		✓		
5	¿Cómo considera usted la identificación de un cliente moroso?	✓		✓		✓		
6	¿Cómo considera usted la toma de decisiones frente a la información del cliente que ha sido evaluado?	✓		✓		✓		
7	¿Cómo considera usted las actividades diarias que se realiza en el departamento de cobranza?	✓		✓		✓		
DIMENSIÓN 3: Tiempo		Si	No	Si	No	Si	No	
8	Tiempo para predecir	✓		✓		✓		

Observaciones (precisar si hay suficiencia): Si hay suficiencia. Puede aplicarse

Opinión de aplicabilidad: **Aplicable [✓]** **Aplicable después de corregir [ ]** **No aplicable [ ]**

Apellidos y nombres del juez validador. Dr/ Mg: Mg. Luis Cuya **DNI:** 09553506

Especialidad del validador: Mg. Informática

Cel: 997820072

03 de 12 del 20 18

Firma del Experto Informante.

[1]**Pertinencia:** El ítem corresponde al concepto teórico formulado.
[2]**Relevancia:** El ítem es apropiado para representar al componente o dimensión específica del constructo
[3]**Claridad:** Se entiende sin dificultad alguna el enunciado del ítem, es conciso, exacto y directo

Nota: Suficiencia, se dice suficiencia cuando los ítems planteados son suficientes para medir la dimensión

CERTIFICADO DE VALIDEZ DE CONTENIDO DE INSTRUMENTOSA TRAVÉS DE JUICIO DE EXPERTO

Título de la investigación	**Aplicación de minería de datos basado en árboles de decisión para predecir el riesgo de morosidad de los clientes en la empresa de seguros Oncosalud S.A.C. 2018**
Nombre(s) del(os) instrumento(s)	**Cuestionario, Ficha de registro**
Autor(es) del instrumento	**Leonardo Espino Quiñones** **Maria Emily Garcia Torres**

Nº	DIMENSIONES / Indicadores	Pertinencia[1]		Relevancia[2]		Claridad[3]		Sugerencias
DIMENSIÓN 1: Precisión		**Si**	**No**	**Si**	**No**	**Si**	**No**	
1	Precisión de predicción	✓		✓		✓		
2	Error de predicción	✓		✓		✓		
DIMENSIÓN 2: Dificultad		**Si**	**No**	**Si**	**No**	**Si**	**No**	
3	¿Cómo considera usted la recolección de la información necesaria para determinar si un cliente podría ser moroso?	✓		✓		✓		
4	¿Cómo considera usted la evaluación de la información del cliente?	✓		✓		✓		
5	¿Cómo considera usted la identificación de un cliente moroso?	✓		✓		✓		
6	¿Cómo considera usted la toma de decisiones frente a la información del cliente que ha sido evaluado?	✓		✓		✓		
7	¿Cómo considera usted las actividades diarias que se realiza en el departamento de cobranza?	✓		✓		✓		
DIMENSIÓN 3: Tiempo		**Si**	**No**	**Si**	**No**	**Si**	**No**	
8	Tiempo para predecir	✓		✓		✓		

Observaciones (precisar si hay suficiencia): .....

Opinión de aplicabilidad: **Aplicable [X]** **Aplicable después de corregir [ ]** **No aplicable [ ]**

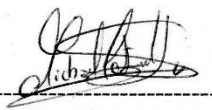
Apellidos y nombres del juez validador. Dr/ Mg: Cabanillas Carbonell Michael **DNI:** 43426369

Especialidad del validador: Ing. Sistemas

Cel: 987807040

[1]**Pertinencia:** El ítem corresponde al concepto teórico formulado.
[2]**Relevancia:** El ítem es apropiado para representar al componente o dimensión específica del constructo
[3]**Claridad:** Se entiende sin dificultad alguna el enunciado del ítem, es conciso, exacto y directo

Nota: Suficiencia, se dice suficiencia cuando los ítems planteados son suficientes para medir la dimensión



Firma del Experto Informante.

APÉNDICE V: Cuestionario de la Pre-prueba

CUESTIONARIO

El presente cuestionario tiene como objetivo identificar el nivel de dificultad de los empleados del departamento de cobranza respecto a la predicción del riesgo de morosidad de los clientes de la empresa Oncosalud.

Este cuestionario consta de 5 preguntas, lea atentamente cada una de ellas y responda.

Marca con un **aspa** o un **check** la opción más apropiada que considere.

Nombre: Juan Morato

1: Muy difícil 2: Difícil 3: Normal 4: Fácil 5: Muy fácil

		1	2	3	4	5
1	¿Cómo considera usted la recolección de la información necesaria para determinar si un cliente podría ser moroso?		/			
2	¿Cómo considera usted la evaluación de la información del cliente?		/			
3	¿Cómo considera usted la identificación de un cliente moroso?		/			
4	¿Cómo considera usted la toma de decisiones frente a la información del cliente que ha sido evaluado?		/			
5	¿Cómo considera usted las actividades diarias que se realiza en el departamento de cobranza?		/			

APÉNDICE VI: Cuestionario de la Post-prueba

CUESTIONARIO

El presente cuestionario tiene como objetivo identificar el nivel de dificultad de los empleados del departamento de cobranza respecto a la predicción del riesgo de morosidad de los clientes de la empresa Oncosalud.

Este cuestionario consta de 5 preguntas, lea atentamente cada una de ellas y responda.

Marca con un **aspa** o un **check** la opción más apropiada que considere.

Nombre: Juan Marcelo

1: Muy difícil 2: Difícil 3: Normal 4: Fácil 5: Muy fácil

		1	2	3	4	5
1	¿Cómo considera usted la recolección de la información necesaria para determinar si un cliente podría ser moroso?					✓
2	¿Cómo considera usted la evaluación de la información del cliente?					✓
3	¿Cómo considera usted la identificación de un cliente moroso?				✓	
4	¿Cómo considera usted la toma de decisiones frente a la información del cliente que ha sido evaluado?					✓
5	¿Cómo considera usted las actividades diarias que se realiza en el departamento de cobranza?					✓

APÉNDICE VII: Carta de conformidad por implementación del sistema web

CARTA DE CONFORMIDAD POR IMPLEMENTACIÓN DE SISTEMA WEB PARA PREDECIR RIESGOS DE MOROSIDAD DE CLIENTES

Lima 10 de diciembre de 2018

Sr.
José Luis Herrera Salazar
Director de Carrera Profesional de Ingeniería de Sistemas
Facultad de Ingeniería
Universidad Autónoma del Perú
Presente. -

De nuestra consideración

Es grato dirigirme a ustedes en representación de ONCOSALUD S.A.C. para hacer de su conocimiento que la señorita Garcia Torres Maria Emily y el señor Espino Quiñones Leonardo, han cumplido a la fecha con implementar el Sistema web para predecir riesgos de morosidad de clientes, como parte del Proyecto de Investigación denominado "Aplicación de minería de datos basado en arboles de decisión para predecir el riesgo de morosidad de los clientes en la empresa de seguros Oncosalud SAC 2018", que fue autorizado para su ejecución y uso de información en nuestra empresa en el mes de diciembre 2017.

Vale la oportunidad, para indicar que la señorita Garcia Torres Maria Emily y el señor Espino Quiñones Leonardo han demostrado responsabilidad en el proyecto mencionado. Es por ello que emitimos esta carta a pedido de los interesados.

Sin otro particular, quedo de usted

Atentamente

ONCOSALUD S.A.C. V°B° ALDO RAMIREZ SILVA JEFE DE RECUPERO

Aldo Ramirez Silva
Jefe de Recupero de clientes