



FACULTAD DE INGENIERÍA Y ARQUITECTURA
ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y SISTEMAS

**PROPUESTA DE MODELO DE DETECCIÓN DE FRAUDES DE
ENERGÍA ELÉCTRICA EN CLIENTES RESIDENCIALES DE LIMA
METROPOLITANA APLICANDO MINERÍA DE DATOS**

PRESENTADA POR

JOHANNA DENISE FLORES COAGUILA

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE
INGENIERO DE COMPUTACIÓN Y SISTEMAS

LIMA – PERÚ

2014



**Reconocimiento - No comercial - Compartir igual
CC BY-NC-SA**

El autor permite transformar (traducir, adaptar o compilar) a partir de esta obra con fines no comerciales, siempre y cuando se reconozca la autoría y las nuevas creaciones estén bajo una licencia con los mismos términos.

<http://creativecommons.org/licenses/by-nc-sa/4.0/>



USMP
UNIVERSIDAD DE
SAN MARTÍN DE PORRES

**FACULTAD DE
INGENIERÍA Y ARQUITECTURA**

**ESCUELA PROFESIONAL DE INGENIERÍA DE COMPUTACIÓN Y
SISTEMAS**

**PROPUESTA DE MODELO DE DETECCIÓN DE FRAUDES DE
ENERGÍA ELÉCTRICA EN CLIENTES RESIDENCIALES DE
LIMA METROPOLITANA APLICANDO MINERÍA DE DATOS**

TESIS

**PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE
COMPUTACIÓN Y SISTEMAS**

PRESENTADO POR

FLORES COAGUILA, JOHANNA DENISE

LIMA – PERÚ

2014

DEDICATORIA

A mis padres, Ana y Virgilio, por su valioso apoyo e incondicional. A mi hermano, Carlos y a mis abuelos, Luisa y Dionisio, por fortalecerme con su amor y aliento.

ÍNDICE

	Página
RESUMEN	X
ABSTRACT	XI
CAPÍTULO I. MARCO TEÓRICO	1
1.1 Antecedentes de la investigación	1
1.2 Bases teóricas	4
1.3 Definición de términos básicos	38
CAPÍTULO II. METODOLOGÍA	41
2.1 Aspectos metodológicos de la investigación	41
2.2 Tipos de Metodología	44
2.3 Recursos	45
2.4 Evaluación de la Viabilidad Económica	48
2.5 Cronograma de Actividades	52
CAPÍTULO III. DESARROLLO DEL PROYECTO	55
3.1. Esquema solución propuesta para el desarrollo del modelo predictivo	55
3.2. Metodología CRISP – DM para la creación y desarrollo del modelo predictivo.	58
CAPÍTULO IV. PRUEBAS Y RESULTADOS	139
CONCLUSIONES	148
RECOMENDACIONES	149
FUENTES DE INFORMACIÓN	150
ANEXOS	156

ÍNDICE DE TABLAS

	Página
Tabla 1: Variables independientes de Consumidores de energía eléctrica.	41
Tabla 2: Variables Independientes de Consumidores fraudulentos	43
Tabla 3: Variable dependiente	44
Tabla 4: Recursos del Proyecto	47
Tabla 5: Hardware y Software del Proyecto	47
Tabla 6: Inversión Inicial del Proyecto	48
Tabla 7: Monto a recuperar con el Proyecto	48
Tabla 8: Ingresos Adicionales del Proyecto	48
Tabla 9: Egresos Adicionales del Proyecto	49
Tabla 10: Flujo Neto del Proyecto	49
Tabla 11: Fuente de Fondos	49
Tabla 12: Valor Actual Neto	50
Tabla 13: Calculo del TIR i2	50
Tabla 14: Calculo del TIR i1	51
Tabla 15: Información Inicial para la Investigación	60
Tabla 16: Estructura de Datos Información Original BaseConsumos	62
Tabla 17: Estructura de Datos de la Información a Investigar Valorizaciones	63
Tabla 18: Tablas a utilizar en el proceso ETL para la Preparación de Datos	75
Tabla 19: Diccionario de Datos de la Tabla BaseConsumoPre	76
Tabla 20: Diccionario de datos de la Tabla ValorizacionPre	78
Tabla 21: Diccionario de Datos de la Tabla BaseConsumos	78
Tabla 22: Diccionario de Datos de la Tabla Valorizacion	80
Tabla 23: Diccionario de Datos de la Tabla Consumos	81
Tabla 24: Diccionario de Datos de la Tabla Maestro	81
Tabla 25: Diccionario de Datos de la Tabla PeriodoActual	82
Tabla 26: Diccionario de Datos de la Tabla FiltroMaestro	83
Tabla 27: Diccionario de Datos de la Tabla Consulta	83

Tabla 28: Descripción de Parámetros del Procedimiento SP_0_CARGA_TOTAL	94
Tabla 29: Descripción de Procedimiento SP_0_Carga_Total	95
Tabla 30: Descripción de Procedimiento SP_0_CARGA_BASES_CONSUMOS	96
Tabla 31: Descripción de Procedimiento SP_0_CARGA_VALORIZACION	98
Tabla 32: Descripción de procedimiento Sp_1_Prepara_Datos	101
Tabla 33: Descripción de Procedimiento SP_2_Carga_Hurto_Maestro	103
Tabla 34: Descripción de Procedimiento sp_3_carga_hurto_casos	105
Tabla 35: Descripción de Matriz de Clasificación	128
Tabla 36: Resultado de Cantidad de Energía Detectada.	138
Tabla 37: Cuadro de Pruebas y Resultados	147

ÍNDICE DE FIGURAS

	Página
Figura 1: Evolución del Indicador % de Pérdidas. Periodo 1994-2011.	xv
Figura 2: Clasificación de Pérdidas de energía	6
Figura 3: Arquitectura de BI	9
Figura 4: Modelos de Minería de Datos.	12
Figura 5: Estructura general de una neurona biológica.	20
Figura 6: Funcionamiento general de una Neurona artificial.	21
Figura 7: Proceso de minería de datos.	23
Figura 8: Medidas de Tendencia Central - La Mediana	25
Figura 9: Medidas de Tendencia Central - La Moda	25
Figura 10: Encuesta en que industria se aplica DM en el 2012	28
Figura 11: Fases de la Metodología SEMMA	29
Figura 12: Esquema de Extracción de los cuatro niveles de abstracción de la metodología CRISP-DM	34
Figura 13: Metodología de Minería de Datos - Fases de la Metodología CRISP-DM	34
Figura 14: Comparativa de las interrelaciones entre las fases de las metodologías SEMMA y CRISP-DM	37
Figura 15: Cronograma de Actividades General	52
Figura 16: Cronograma de Actividades-Parte 1	53
Figura 17: Cronograma de Actividades-Parte 2	54
Figura 18: Esquema solución propuesta para crear el modelo predictivo de detección de fraudes de energía eléctrica en clientes residenciales	56
Figura 19: Pérdidas Técnicas y No Técnicas	59
Figura 20: Información Inicial para la Investigación	61
Figura 21: Campos Iniciales a Investigar	61
Figura 22: Diccionario de Claves de Lectura	65
Figura 23: Histograma frecuencia Códigos de Lectura	65
Figura 24: Histograma frecuencia Claves de Facturación	67

Figura 25: Análisis de clave de facturación	68
Figura 26: Análisis del consumo en kilowatts	68
Figura 27: Histograma muestra consumo promedio logarítmico	69
Figura 28: Frecuencia de días facturados	70
Figura 29: Estructura de variable código de facturación.	72
Figura 30: Días de facturación real y calculada.	73
Figura 31: Ejemplo de cálculo real de los días de facturación.	74
Figura 32: Importación de fuentes de datos originales.	87
Figura 33: Selección de Tipo de Archivo de Origen de Datos.	88
Figura 34: Selección de Origen de Destino de archivos originales.	88
Figura 35: Selección de Base de Datos donde se cargarán las fuentes iniciales.	89
Figura 36: Selección de datos a copiar a la Base de Datos.	89
Figura 37: Selección de tabla destino en la Base de datos.	90
Figura 38: Seleccionar las columnas de origen y destino en la carga coincidan.	90
Figura 39: Validación de carga satisfactoria.	91
Figura 40: Secuencia de procedimiento de preparación de Datos	91
Figura 41: Proceso de Preparación de Datos.	100
Figura 42: Proceso de Carga Hurto a la tabla Maestro	103
Figura 43: Extraer Casos de Hurto	106
Figura 44: Modelo del Perceptrón Multicapa - Retropropagación.	107
Figura 45: Modelo de red neuronal Multicapa.	108
Figura 46: Estructura de Datos para Entrenar Modelo	111
Figura 47: Modelo de Red Neuronal	111
Figura 48: Conexión a Análisis Services.	118
Figura 49: Creación de Base de Datos en Análisis Services.	119
Figura 50: Probar Conexión con a la BD con el Complemento de Minería de Datos.	119
Figura 51: Conexión a la BD de Análisis Services.	120
Figura 52: Creación de Origen de Datos - Clasificar Datos.	120
Figura 53: Crear un Nuevo Origen de Datos.	121
Figura 54: Datos del Servidor de Análisis Services para crear origen de datos.	122
Figura 55: Creación de Modelo – Clasificar.	122
Figura 56: Creación de Modelo - Asistente para Clasificar	123
Figura 57: Creación de Modelo - Selección de Datos de Entrenamiento.	123
Figura 58: Creación de Modelo - Selección de columna a Analizar.	124
Figura 59: Creación de Modelo - Selección de Algoritmo y parámetros de la Red Neuronal.	124
Figura 60: Creación de Modelo - Selección de Datos de Prueba.	125
Figura 61: Creación de Modelo - Nombre de Estructura y Modelo.	125
Figura 62: Creación de Matriz de Clasificación.	126
Figura 63: Matriz de Clasificación - Selección de Estructura y Modelo.	126

Figura 64: Matriz de Clasificación - Seleccionar columna que se va a predecir.	127
Figura 65: Matriz de Clasificación - Selección de Origen de Datos y Vista.	127
Figura 66: Resultado de la Matriz de Clasificación.	128
Figura 67: Evaluación de Modelo - Consulta.	129
Figura 68: Evaluación de Modelo - Selección de Modelo a Consultar.	130
Figura 69: Evaluación de Modelo - Selección de Datos a Evaluar.	131
Figura 70: Evaluación de Modelo- Relación de columnas del modelo con la data a consultar. Elaboración: La autora	131
Figura 71: Evaluación de Modelo - Agregar Salida Número de Cuenta del Cliente.	132
Figura 72: Evaluación de Modelo - Agregar Salida Hurto.	132
Figura 73: Evaluación del Modelo - Agregar Salida Sospecha.	133
Figura 74: Evaluación de Modelo - Resultado de Evaluación.	133
Figura 75: Resultado de Consulta del Modelo	134
Figura 76: Nuevos datos para Entrenamiento.	135
Figura 77: Datos a evaluar sospechoso.	136
Figura 78: Complemento Minería de Datos – Opción Consulta	136
Figura 79: Complemento Minería de Datos – Opción Asistente de Consulta	136
Figura 80: Complemento Minería de Datos – Opción Datos Origen	137
Figura 81: Complemento Minería de Datos – Opción relación de columnas	137
Figura 82: Complemento Minería de Datos – Resultado de Consulta	138
Figura 83: Resultado de Prueba 01	141
Figura 84: Resultado de Prueba 02	142
Figura 85: Resultado de Prueba 03	144
Figura 86: Resultado de Prueba 04	146
Figura 87: Sustentación Tesis Pag. 01	156
Figura 88: Sustentación Tesis Pág. 02	157
Figura 89: Sustentación Tesis Pág. 03	157
Figura 90: Sustentación Tesis Pág. 04	158
Figura 91: Sustentación Tesis Pág. 05	158
Figura 92: Sustentación Tesis Pág. 06	159
Figura 93: Sustentación Tesis Pág. 07	159
Figura 94: Sustentación Tesis Pág. 08	160
Figura 95: Sustentación Tesis Pág. 09	160
Figura 96: Sustentación Tesis Pág. 10	161
Figura 97: Sustentación Tesis Pág. 11	161
Figura 98: Sustentación Tesis Pág. 12	162
Figura 99: Sustentación Tesis Pág. 13	162
Figura 100: Sustentación Tesis Pág. 14	163
Figura 101: Sustentación Tesis Pág. 15	163

Figura 102: Sustentación Tesis Pág. 16	164
Figura 103: Sustentación Tesis Pág. 17	164
Figura 104: Sustentación Tesis Pág. 18	165
Figura 105: Sustentación Tesis Pág. 19	165

RESUMEN

En el presente trabajo, desarrollaremos un modelo como propuesta para predecir potenciales situaciones de fraudes de energía eléctrica en clientes residenciales basado en aprender el comportamiento de clientes que anteriormente hurtaron para ello aplicaremos el proceso Minería de Datos basado para analizar, extraer y almacenar información de la base de datos, las cual contiene el historia del consumos de energía.

El modelo se propone para apoyar a las empresas de distribución eléctricas en especial a los técnicos eléctricos a examinar y verificar, acertadamente, de manera rápida y oportuna los resultados obtenidos y contribuya, de esta forma, en la toma de decisiones. Para la creación del modelo se utilizaron las redes neuronales por ser la de mejor desempeño en la detección. El modelo creado fue evaluado con datos de una empresa distribuidora de electricidad para el período 2009 y 2010. Las herramientas utilizadas para la creación del modelo fueron el Sql Server Management Studio (Database Engine para la base de datos y creación de procedimientos, Análisis Services para la creación de Estructuras y modelos) y como herramienta interactiva y de fácil entendimiento para el usuario el Complemento de Minería de Datos para Microsoft Excel. Ésta última, el usuario la utilizará para crear, evaluar y consultar los modelos ya existentes o nuevos modelos.

Palabras Claves: Minería de Datos, Redes Neuronales, Fraude de energía.

ABSTRACT

In this work a model is developed as a proposal for predicting potential fraud situations of electric energy in residential customers based on learning clients' behaviour who stole it previously, therefore we are going to apply the Data Mining process based on analyzing, extracting and storing database information, which includes the record of energy consumption.

The purpose of this model is to support electric distribution enterprises, especially for electrical technicians to examine and verify, accurately, quick and suitable manner the results obtained and to contribute in this way, on decision making. For model creation, neural networks were used for being the best performers in the detection. The created model was evaluated using data from an electricity distributor enterprise for the period of 2009 and 2010. The tools used to create the model were the SQL Server Management Studio (Database Engine for database and creating procedures, Analysis Services for designing structures and models) and as an interactive tool and easy to understand for the user, Data Mining Add-in for Microsoft Excel. This latter, the user will employ to create, evaluate and consult existing models or new models.

Keywords: Data mining, neural networks, Energy fraud.

INTRODUCCIÓN

La problemática principal de las compañías de electricidad son las pérdidas de energía eléctrica por parte de los consumidores que utilizan diversos métodos como toma clandestina y alteración del funcionamiento de medidores para el pago real de los consumos realizados. Esto lo evidencia la empresa Edelnor donde publica en la memoria anual 2011 el porcentaje de pérdidas de energía eléctrica llega alcanzar el 8.16% de la energía comercializada.

Actualmente las empresas no cuentan con procedimientos automatizados de reducción de pérdidas de energía eléctrica eficientes debido a que día a día la demanda de energía aumenta y las modalidades de fraude cambian constantemente una vez encontradas, los consumidores deshonestos perfeccionan sus prácticas ilegales, y a pesar de los avances tecnológicos en el campo de la medición.

Por tal motivo, es necesario contar con un modelo automatizado que permita detectar los fraudes de energía eléctrica de manera rápida y oportuna, y así ayude a predecir a los consumidores sospechosos de hurto.

El Modelo creado se propone para detectar los fraudes de energía eléctrica no solo contribuye con la empresa distribuidora con el recupero ingresos y disminución de gastos, sino también con la población más necesitada de la sociedad, ya que esta energía recuperada se distribuya acertadamente en la población más necesitada a la sociedad.

El problema de las empresas distribuidoras de energía eléctricas es que no cuentan con un modelo automatizado que apoye en la detección de energía eléctrica en clientes residenciales, debido a que las modalidades de hurto cambian constantemente ya que los consumidores deshonestos perfeccionan sus prácticas una vez detectados, la demanda de energía aumenta y se necesita tratar mayores volúmenes de datos y el proceso para detectar la energía toma mayor tiempo.

La empresa distribuidora Edelnor cuenta con un área de recupero de energía eléctrica y año a año logra reducir el porcentaje de pérdidas de energía eléctrica, para el año 2011 es el 8.16% de la energía que comercializa esto lo evidencia en la memoria anual 2011(Capítulo Nuestra Energía, Control de Pérdidas p23), donde indica que de 193,516 inspecciones logró recuperar 17419 consumos no registrados esto equivaldría a un 9% porcentaje de éxito, el recupero en mención se debió a proyectos de reinserción de clientes, compra de 160 medidores inteligentes de alta tecnología.

Como problema, se formuló ¿De qué modo se puede detectar clientes fraudulentos para minimizar las pérdidas de energía eléctrica?

El objetivo general es detectar fraudes de energía eléctrica mediante un modelo basado en aprender comportamientos de clientes que anteriormente hurtaron y permita predecir clientes sospechosos de hurto. Entre los objetivos específicos, se precisan desarrollar un modelo para la detección de fraudes de energía eléctrica en clientes residenciales utilizando

Minería de Datos, evaluar el modelo para la detección de fraudes de energía mejorando su rendimiento, probar y afinar el modelo desarrollado.

Como impacto de las pérdidas no técnicas se sostiene que las empresas distribuidoras de energía eléctrica se ven muy perjudicadas por las Pérdidas no- técnicas como indicaron Ríos, Uribe (2013) que debido al fraude, imprecisión de la medición, errores de lectura o mala gestión de los sistemas de recogida de datos, las compañías eléctricas pierden grandes cantidades de dinero cada año debido al fraude por consumidores de electricidad y es difícil diferenciar entre los clientes honestos y fraudulentos. (p. 18)

Ríos, Uribe (2013) también indicaron que las empresas de distribución eléctrica evalúan el impacto en las pérdidas técnicas que se dan en las redes de generación, transmisión y distribución midiendo el rendimiento global y pudiendo ser calculada, pero el caso de las pérdidas no técnicas, que son consecuencia del hurto de energía electricidad por manipulación del medidor, medición defectuosa y errores de facturación, son las de mayor preocupación en todo el mundo debido a que no se puede calcular, prevenir y detectar tempranamente.

Las pérdidas que enfrentan las empresas de distribución de energía eléctrica como Edelnor indican, todos los años, un porcentaje indicador de pérdidas de energía que se redujo a 8.16% en el periodo 2011, donde muestran que de 193 516 inspecciones se logró recuperar 17419 conexiones por concepto de consumos no registrados, equivalentes a S/. 6.91 millones de nuevos soles, de lo que podemos deducir que el porcentaje de éxito fue el 9%. Este recupero de energía se debió a una inversión de 160 medidores inteligentes que identifican los alimentadores con el mayor índice de pérdidas y se despliega personal técnico dedicado a la detección de pérdidas y se despliega personal técnico dedicado a la detección de pérdidas y con proyectos de reinserción de clientes donde se propuso nuevas conexiones y reducir el consumo no registrado en los conos y comunidades en situaciones de escasos recursos económicos.

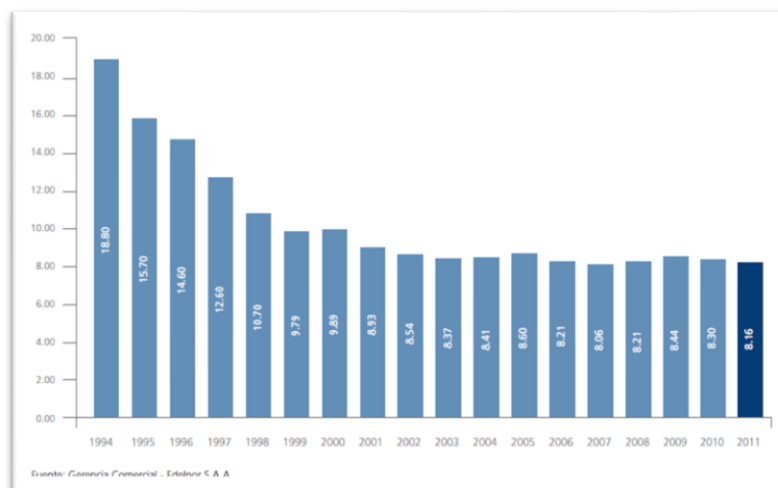


Figura 1: Evolución del Indicador % de Pérdidas. Periodo 1994-2011.

Elaboración: La autora

La presente tesis se justifica por las siguientes razones. Justificación Metodológica debido a que el resultado de la presente tesis es crear un modelo automatizado utilizando minería de datos, donde indicaremos los procesos, pasos y tratamientos que se deben seguir los datos originales para la detección de fraudes de energía eléctrica con un alto porcentaje de probabilidad de fraude de manera ágil, oportuna, y el resultado obtenido apoya a la toma de decisiones e investigaciones futuras.

El crear un modelo automatizado que detecte casos de fraudes de energía eléctrica, permitirá que, con mayor eficiencia, se planifique la inspección de suministros que detecte las condiciones fraudulentas, mejorando la productividad de las empresas de servicio eléctrico, lo cual tendrá un impacto positivo en la sustentabilidad financiera y en la calidad del servicio.

Como justificación práctica el modelo de detección de fraudes de energía eléctrica tiene un alto porcentaje de especificidad y sensibilidad. Además, ayudará al área de recupero energía eléctrica a que mejorar el proceso de detección de pérdidas no técnicas, ya que brindaremos una herramienta de consulta que permitirá procesar grandes volúmenes de datos, extraer y aprender comportamientos en un corto tiempo y evaluarlo,

si corresponde a usuarios fraudulentos. El resultado del proyecto permitirá el recupero de energía, en menor tiempo y menores costos en alta tecnología de medidores inteligentes y cantidad de personal de campo y así evitar subjetividad.

La justificación social consiste en que el servicio eléctrico es considerado de primera de necesidad, en este sentido su calidad y sustentabilidad de las empresas encargadas de este impactan, positivamente, en la calidad de vida de los usuarios y de la sociedad en general. Las condiciones fraudulentas atentan contra la calidad del servicio de usuarios legalmente conectados.

La propuesta de crear un modelo automatizado para detectar las pérdidas de energía eléctrica permitirá mejorar la calidad de vida de las personas que no tienen acceso a la energía eléctrica debido a la mala utilización y distribución de la misma por usuarios fraudulentos.

La detección de pérdidas no técnicas de energía eléctrica evitará el riesgo de exponer vidas por la manipulación indebida e inadecuada de cables, medidores y demás instrumentos que de distribución de electricidad.

CAPÍTULO I

MARCO TEÓRICO

1.1 Antecedentes de la investigación

A. Proyecto de Minería de datos aplicada a la detección de clientes con alta probabilidad de fraudes en sistemas de distribución.

Ríos y Uribe (2013) indicaron que, en este proyecto se emplearon máquina de soporte vectorial (SVM) para la clasificación de datos y elección con una alta probabilidad estos datos pueden ser importantes para la decisión que se encuentra en proceso de análisis. Esta investigación utiliza como fuente de información perfiles de carga de usuarios residenciales con el fin de construir y caracterizar dichos comportamientos, por medio de patrones de consumo reales que se pueden presentar para un periodo de tiempo determinado y permitir su representación cuantificada, por medio de herramientas computacionales que definen su estado (normal o sospechoso). Es aquí donde la SVM se encarga de utilizar los perfiles de carga y clasificarlos según sus comportamientos. Las SVMs se basan, principalmente, en aprendizajes supervisados que requieren de etapas previas de entrenamiento, y es aquí donde se buscan modelos de clasificación más eficientes y confiables, que garanticen una alta precisión sin importar el número de perfiles de carga que se desea estudiar. (p. 2)

B. Modelo de detección de fraude basado en el descubrimiento simbólico de reglas de clasificación extraídas de una red neuronal.

Santamaría (2010) indicó que el proyecto desarrollado presenta un modelo de detección de fraude empleando técnicas de minería de datos tales como redes neuronales y extracción simbólica de reglas de clasificación a partir de la red neuronal entrenada. La propuesta del modelo surge del interés de diseñar y desarrollar una herramienta de detección de fraude, con el fin de ayudar a expertos del negocio a examinar y verificar fácilmente los resultados obtenidos para apoyar la toma de decisiones.

El modelo propuesto se probó con un conjunto de datos de una organización colombiana para el envío y pago de remesas, con el fin de identificar patrones ligados a la detección de fraude. De igual forma, los resultados de las técnicas utilizadas en el modelo, se compararon con otras técnicas de minería como los árboles de decisión. Para ello, un prototipo de software se desarrolló para probar el modelo, que fue integrado a la herramienta RapidMiner y que puede ser usado como una herramienta de software académico. (p. 12)

C. Detección de consumos anómalos de energía eléctrica utilizando nuevas características

Rodríguez y Castro (2012) indicaron que el problema de detección de clientes fraudulentos es, en general, de gran importancia para empresas proveedores de diversos rubros. En vista de esta problemática, se propone investigar la utilidad de incluir a la herramienta de detección de consumos anómalos (DeCA) nuevas características. Para esto, se propone emplear selección del conjunto formado por las características consideradas en trabajos anteriores y nuevas características proporcionadas por técnicos de UTE. Esta base consta de 114 registros etiquetados como sospechosos y 679 etiquetados como anómalos (p. 4).

D. Estrategia inteligente eficiente para la planificación de las inspecciones de suministros de las empresas del servicio eléctrico.

Lima (2011) indicó que las pérdidas fraudulentas generan, en las empresas de servicio eléctrico y en la sociedad, efectos que impactan negativamente. Debido a que produce fallas del servicio e incremento de pérdidas económicas para las empresas. (Las empresas están tomando diversas acciones para prevenir y controlar la condición de fraude. La estrategia inteligente propuesta además de hacer minería de datos, es decir, extraer de la data información importante y escondida de los consumos de los usuarios, también utilizaría información de los expertos; es decir, que incluiría un sistema experto además de la base de conocimientos. En este sentido, se recomiendan incorporar conocimiento de expertos, más complejos y diferentes parámetros a los ya aplicados para el diseño de redes neuronales así como otras técnicas de minería de datos.

La eficiencia de la aplicación de la estrategia inteligente para la planificación de inspecciones a usuarios de las empresas de servicio eléctrico se comparará con las obtenidas utilizando otras herramientas. El diseño y validación de la estrategia inteligente se propone dirigirla al sector residencial, ya que el número de usuarios de este es mayor que el de los sectores industriales y comerciales, característica que lo hace contribuir de manera significativa en la disminución de las pérdidas por condiciones fraudulentas. (p. 17)

Caso de estudio

Caso de éxito de Aplicación de Minería de datos para la detección de Anomalías utilizando Metodología CRISP – DM Cravero, Sepúlveda (2009) explicaron que la Minería de Datos es una de las soluciones de la Inteligencia de Negocios, que ayuda a extraer conocimiento a partir de datos que las empresas han generado producto de su negocio. Aplicaciones tales como Selling, optimización de cadena de suministros o conceptos tales como clasificadores y regresiones, basadas en redes neuronales han

emergido profundamente durante los últimos años. En este contexto, se implementó una aplicación de Minería de Datos que detecta anomalías con el fin de detectar posibles abusos en el uso de los principales servicios que otorgan las empresas de agua. Los datos corresponden a las facturaciones de servicios de agua potable y alcantarillado de clientes en la ciudad de Lautaro. Para el análisis, se utilizaron algoritmos de Minería de Datos, basados en técnicas de Clustering y la metodología CRISP-DM. Como principal resultado, el sistema permitió a la empresa reducir de manera considerable el tiempo de búsqueda de los posibles fraudes (p. 1).

1.2 Bases teóricas

1.2.1 Pérdidas de energía

Ríos y Uribe (2013) afirman que las pérdidas de potencia o pérdida de energía, se definen como la diferencia entre la energía suministrada o entregada y la energía facturada o consumida por el cliente, como se ilustra en la siguiente ecuación:

$$E_{\text{Loss}} = E_{\text{delivered}} - E_{\text{SOld}}$$

Donde:

E_{Loss} : Es la cantidad de energía perdida

$E_{\text{delivered}}$: Representa la cantidad de energía suministrada.

E_{SOld} : Representa la cantidad de energía registrada o vendida. (p.14).

Clasificación de pérdidas de energía eléctrica:

Las pérdidas de energía se pueden dividir en dos áreas principales que son: pérdidas técnicas y pérdidas no técnicas.

A. Pérdidas técnicas

Ríos y Uribe (2013) mencionan que las pérdidas técnicas se producen durante la transmisión y distribución, e involucran a la subestación, transformador y las pérdidas en las líneas de distribución. Estos incluyen las pérdidas resistivas en los alimentadores primarios de distribución, pérdidas al interior del transformador (pérdidas resistivas en los devanados y pérdidas en el núcleo), las pérdidas resistivas en las redes secundarias y las pérdidas en la medición de energía. Por lo tanto,

$$EL_{oss} = U_{ElectricityCost} \times E_{Loss} + M_{MaintenanceCost}$$

Donde:

EL_{oss} : La pérdida de ingreso debido a pérdidas técnicas.

$U_{ElectricityCost}$: Representa el costo unitario de electricidad

E_{Loss} : Representa la cantidad de energía perdida.

$M_{Maintenance Cost}$: Representa el mantenimiento y gastos

Operacionales. (p.15)

Dicho de otra manera, se dan en los elementos y equipos eléctricos que conforman la red y la medición se controlan con buenas prácticas operativas o con procedimientos de diseño automatizado.

B. Pérdidas no técnicas

Ríos y Uribe (2013) indicaron que las pérdidas no técnicas están relacionadas principalmente con el robo de electricidad y procesos administrativos defectuosos. Estos medios incluyen la manipulación de medidores para grabar menores tasas de consumo, la organización de lecturas falsas por medio de sobornos, o conexiones ilegales, donde dicha conexión no pasa por el medidor y la organización irregular de facturas por parte de empleados de entidades comercializadoras de energía, extendiendo facturas a más bajo costo, el ajuste de la posición del punto decimal en las facturas, o simplemente, haciendo caso omiso de las facturas pendientes de pago.

$$C_{NNTL} = C_{LOSS} - C_{TechnicalLoss}$$

C_{NNTL} : Es la componente de costo NTL

C_{LOSS} : Es la pérdida de ingreso debido a pérdidas técnicas.

$C_{TechnicalLoss}$: Representa el componente de costo de pérdida técnicas.
(p.15).

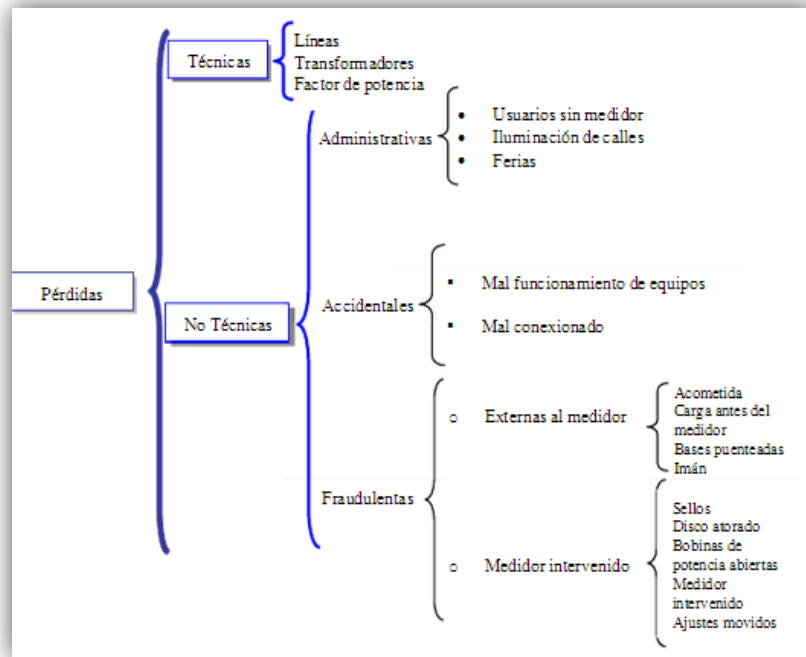


Figura 2: Clasificación de Pérdidas de energía

Fuente: Rodríguez, R. y Gilberto Vidrio “Detección de robos de energía eléctrica”.

Para el caso de la presente tesis, centraremos el estudio de las pérdidas no técnicas.

Clasificación de pérdidas no técnicas:

No toda la energía que se produce, se vende y se factura. Por lo tanto, las empresas suministradoras del servicio de electricidad registran pérdidas en la energía que generan y tiene disponible para su venta. Es decir, una proporción de energía se queda ahí. Los aparatos de medición no los contabilizan como entregados a los usuarios y por lo tanto, no puede ser objetivos de cobro.

Las pérdidas no técnicas no constituyen una pérdida real de energía, esta es utilizada por algún usuario que es suscriptor o no, de la empresa distribuidora la misma que solo recibe parte o ninguna retribución por la prestación del servicio.

Según afirma Cañar (2007), las pérdidas se pueden clasificar de acuerdo con diferentes criterios que los siguientes:

- Clasificación por la causa que los produce.
- Clasificación por su relación con las actividades administrativas de la Empresa.

Clasificación según su causa es:

- a) Por consumo de usuarios no suscritos o contrabando.

Establecida a los usuarios que toman la energía directamente de la red sin entrar en el marco legal de la empresa, por lo que esta energía es consumida, pero no facturada.

- b) Por Error en la contabilización de energía

Está en la lectura, medición y facturación de la energía, esto generalmente se debe a la no simultaneidad de la medición de los contadores.

- Errores en consumo estimado

Son aquellos en que la empresa realiza una estimación de consumo, por ejemplo en caso de abonados temporales.

- Fraude o hurto

Son aquellos que incluidos en la base de dato de los abonados de la empresa, han adulterado los contadores o toman la energía directamente de la red.

- Errores en consumo propio de la empresa.

Corresponde a la energía no contabilizada y consumida por la empresa distribuidora. (p.55).

1.2.2 Técnicas de detección de fraude

La detección de fraude no es un tema trivial, las metodologías usadas por los “falsificadores” no son las mismas de hace algunos años; cuando las entidades identifican un patrón de comportamiento, los “falsificadores” ya están pensando en otras alternativas.

Según indicó Santa María (2010), las herramientas para detección de fraude se pueden clasificar en dos categorías:

- **Técnicas tradicionales**

Los métodos tradicionales de detección de fraude consisten en una combinación de investigadores y herramientas que reportan alarmas de posibles sospechosos; para ello se utilizan técnicas como:

Identificación de clientes que coinciden en listas de control como: OFAC1, BOE2, DATA CREDITO, etc., emitidas por entes internacionales o nacionales.

Otra herramienta son los sistemas basados en la aplicación de reglas que constan de sentencias SQL, definidas con la ayuda de expertos. Esta estructura puede detectar sumas acumulativas de dinero, ingresadas a una cuenta, en un corto periodo de tiempo, como un día.

Por último, métodos de clasificación estadísticos como el análisis de regresión de datos, para detectar comportamientos anómalos de cambio en una cuenta, dada una serie de transacciones que efectúa un cliente en un lapso de tiempo.

1.2.3 Inteligencia de negocios

La Minería de datos es parte de la inteligencia de negocios que apoya a la parte estratégica en la toma de decisiones, brindando conocimiento de que sucederá en base al análisis de datos de hechos históricos.

- La inteligencia de negocios nace como una necesidad de entender no solo que está pasando, sino CUÁNDO, DÓNDE, QUIÉN y POR QUÉ.
- Dar solución a los requerimientos de información de manera oportuna.

Según el instituto de Datos de Datawarehousing, se afirma:

“Business Intelligence se refiere al proceso de convertir datos en conocimiento y conocimiento en acciones para crear la ventaja competitiva del negocio”

Según Deloitte & Touche indica que:

“Los sistemas de Business Intelligence se basan en la integración y en la universalización de la información para dar respuesta a las necesidades de las empresas”.

Arquitectura de BI

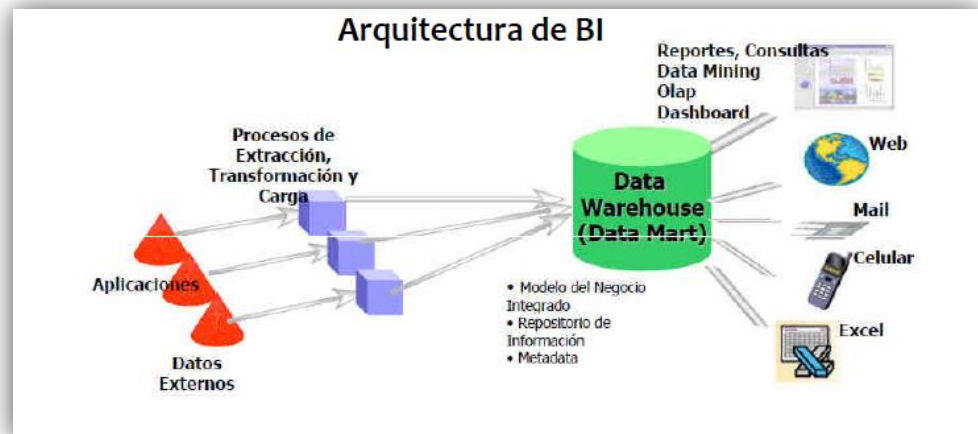


Figura 3: Arquitectura de BI

Fuente: Kasperu(2010)

Datawarehouse

Es un repositorio central de datos, que concentra la información de interés de toda la organización y distribuye dicha información por medio de diversas herramientas de consulta y de creación de informes orientadas a la toma de decisiones.

Los datos almacenados son extraídos de los diferentes sistemas operacionales y fuentes externas.

Datawarehousing

Es el proceso de extraer y filtrar datos de las bases de datos de las operaciones comunes de la empresa, procedentes de los distintos sistemas operacionales, para transformarlos, integrarlos, sumarizarlos y almacenarlos en un depósito o repositorio para poder acceder a ellos cada vez que se necesite.

1.2.4 Minería de datos

Es parte del área de inteligencia de negocio, da un paso más y apoya no solo a entender que pasó, como está y por qué, sino que responde a preguntas como qué pasará y por qué.

Debido a la creciente importancia que se ha dado a la minería de datos, esta ha sido señalada por el Instituto Tecnológico de Massachussets (MIT, 2001) como “Una de las diez tecnologías emergentes más importantes del siglo 21 que cambiará el sentido de la investigación en el mundo”. En efecto, una realidad actual de la minería de datos es su papel de apoyo para explorar, analizar, comprender y aplicar el conocimiento adquirido de grandes volúmenes de datos, así como identificación de tendencias que sirven de ayuda para la toma de decisiones.

En la literatura, se cuenta con varias definiciones para la minería de datos, algunas de ellas son la mencionadas por los autores Ríos et al. (2013) donde postulan que: “La minería de datos es una combinación de bases de datos y tecnologías de inteligencia artificial, también lo definen como el proceso de planteamiento de búsqueda y extracción de patrones, a menudo desconocido de grandes cantidades de datos mediante comparación de patrones y otras técnicas de razonamiento”.(p.5), a esta definición apoya la empresa KASPERU (2010) indicando que la Minería de Datos es el proceso de descubrir conocimiento desde los datos, mediante un proceso de extracción no trivial de información implícita, previamente desconocida y potencialmente útil. También indica que es el conjunto de técnicas para el análisis de los datos y el descubrimiento de patrones

escondidos y concluye indicando que es el conocimiento representado mediante patrones o modelos.

En general, se pueden citar algunos factores que asocian el impulso del empleo de minería de datos según indica Larose (2005):

- El crecimiento exponencial de la recolección de datos y la evolución del poder de cómputo.
- El almacenamiento de los datos en data warehouses, con la finalidad de tener acceso a una base de datos, actualizada y confiable.
- La creciente disponibilidad de información en internet.
- La presión competitiva del mercado en una economía globalizada.
- El desarrollo de herramientas comerciales para llevar a cabo la minería de datos (p.5).

1.2.5 Qué es un Modelo:

Según Kasperu (2010), es un modelo es un intento de entender algún aspecto de la realidad. También indica es un Conocimiento, un modelo. Que intenta representar algún aspecto del mundo y explicar su comportamiento, que permite pasar de la observación a la teoría; se construyen para ser transmitido.

Modelos y tareas de la Minería de Datos

Los modelos empleados, en la minería de datos, son el descriptivo y el predictivo:

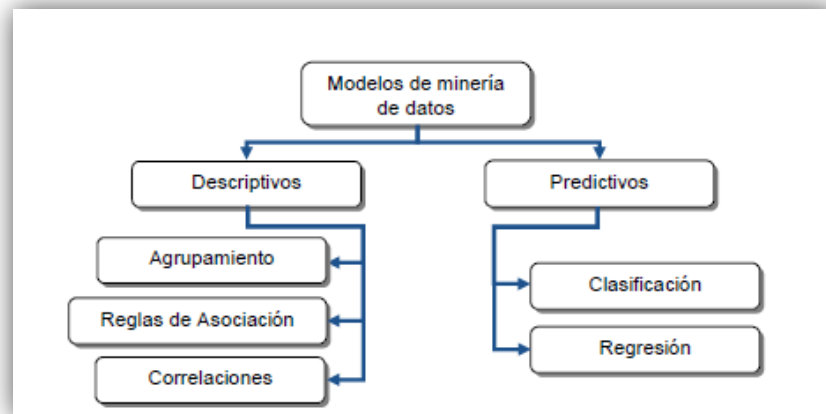


Figura 4: Modelos de Minería de Datos.

Fuente: Calderón (2006)

1.2.5.1 Modelo descriptivo

En el modelo descriptivo, se identifican patrones que describen los datos mediante tarea, agrupamiento y reglas de asociación.

Hernández, Ramírez y Ferri (2004) desatacaron que mediante este modelo se identifican patrones que explican o resumen el conjunto de datos, siendo estos útiles para explorar las propiedades de los datos examinados (p.24).

1.2.5.2 Modelo predictivo

Pretenden estimar valores futuros o desconocidos de variables de interés. Hernández et al. (2004) apoya este postulado y señala que el proceso que se basa en la información histórica de los datos, mediante las cuales se predice un comportamiento de los datos, ya sean mediante clasificaciones, categorizaciones o regresiones (p.25).

1.2.5.3 Tareas de la minería de datos

Dentro de los modelos descriptivos y predictivos se encuentran diferentes tareas específicas como agrupamiento, reglas de asociación, clasificación, regresión, entre otras. En la figura 4, extraída de la tesis de Molero (2008), se muestra una representación general de los modelos y tareas en el proceso de minería de datos:

- **Agrupamiento (Clustering)**

En esta tarea según Berry y Linoff (2004), se evalúan similitudes entre los datos para construir modelos descriptivos, analizar correlaciones entre las variables o representar un conjunto de datos en un pequeño número de regiones.

Consideran al agrupamiento como la tarea de dividir una población heterogénea en un número de subgrupos homogéneos de acuerdo con las similitudes de sus registros. (p.25).

Dentro de esta tarea, existen dos tipos principales de agrupamiento, según indicó Larose (2005): el jerárquico que se caracteriza por el desarrollo recursivo de una estructura en forma de árbol y el particional que organiza los registros dentro de K grupos. (p.33).

En resumen, la agrupación simplemente agrupa los datos según las reglas y patrones inherentes en los datos que se basan en la similitud de sus atributos.

- **Asociación**

Hernández et al. (2004) afirma que mediante esta tarea se identifican afinidades entre la colección de los registros examinados, buscando relaciones o asociaciones entre ellos.

El interés por esta tarea se debe principalmente a que las reglas proporcionan una forma concisa de declarar la información potencialmente útil.

Las ventajas más frecuentes en las reglas de asociación son el descubrimiento de asociación (relaciones que aparecen conjuntamente) y de secuencia (asociada al tiempo) (p .27).

Las reglas de asociación en la minería de datos se utilizan para encontrar hechos que ocurren en común dentro de un conjunto de datos. Dicho de otra manera que deben ocurrir ciertas condiciones para que se produzca cierta condición.

- **Correlaciones**

Las correlaciones son una tarea descriptiva que se usa para determinar el grado de similitud de los valores de dos variables numéricas.

Hernández et al. (2004) añaden que “Una manera de visualizar la posible correlación entre la observación de dos variables (X e Y), es a través de una diagrama de dispersión, en el cual los valores que toman estas variables son representados por puntos. Su principal desventaja es que no puede ser usado para hacer predicciones, puesto que no es clara la forma en que toma la relación” (p.29).

- **Clasificación**

Es una de las principales tareas en el proceso de minería de datos. Calderón (2006) afirma que la clasificación de los datos sirve para identificar las características que indican el grupo al que cada caso pertenezca. Este modelo puede ser usado para comprender los datos existentes y para predecir cómo se comportará algún nuevo caso. Sin embargo, el mayor problema de la clasificación es que muchas veces no es representativo y no proporciona un conocimiento detallado, solo predicciones. (p.28).

Algunos algoritmos comprendidos, en esta fase, son clasificación bayesiana, arboles de decisión, redes neuronales artificiales, entre otros.

- **Regresión**

Para definir esa tarea, nos basamos en los apuntes de Calderón (2006), quien señala que la regresión, utiliza valores existentes para pronosticar qué valores son los que se obtendrán más adelante. En un caso simple de regresión se utilizan técnicas estadísticas como la regresión lineal, desafortunadamente muchos problemas de la vida real no son simples proyecciones lineales de los valores previos. Por ejemplo, los rangos de fallo en volúmenes de ventas de un determinado stock de productos son bastante difíciles de predecir porque dependen de la interacción de múltiples variables de predicción (p.28).

En la regresión según indica Larose (2005), la información de salida es un valor numérico continuo o un vector con valores no discretos (p.51).

Esta constituye la principal diferencia respecto a la clasificación, cuyo valor a predecir es numérico.

1.2.6 Técnicas de minería de datos

1.2.6.1 Redes neuronales

Para introducirnos en este tema citamos al autor Basogain (2008) quien señala que el cerebro humano es el sistema de cálculo más complejo que conoce el hombre. El ordenador y el hombre realizan muy diferentes clases de tareas; así la operación de reconocer el rostro de una persona resulta una tarea relativamente sencilla para el hombre y difícil para el ordenador, mientras que la contabilidad de una empresa es tarea costosa para un experto contable y una sencilla rutina para un ordenador básico.

La capacidad del cerebro humano de pensar, recordar y resolver problemas ha inspirado a muchos científicos a intentar o procurar modelar en el ordenador el funcionamiento del cerebro humano. (p.1).

Otras definición que se puede rescatar de la cátedra: Información aplicada a la ingeniería de procesos, indicados por el autor Matich (2001), son las siguientes:

Una nueva forma de computación, inspirada en modelos biológicos.

Un modelo matemático compuesto por un gran número de elementos procesales organizados en niveles.

Un sistema de computación compuesto por un gran número de elementos simples, elementos de procesos muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas.

Redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico. (p.8)

1.2.6.2 Características de las redes neuronales artificiales

Basogain (2008) señaló que las Redes Neuronales Artificiales, ANN (Artificial Neural Networks) están inspiradas en las redes neuronales biológicas del cerebro humano. Están constituidas por elementos que se comportan, de forma similar a la neurona biológica, en sus funciones más comunes. Estos elementos están organizados de una forma parecida a la que presenta el cerebro humano.

- **Aprender**

Adquirir el conocimiento de una cosa por medio del estudio, ejercicio o experiencia. Las ANN pueden cambiar su comportamiento en función del entorno. Se les muestra un conjunto de entradas y ellas mismas se ajustan para producir unas salidas consistentes.

- **Generalizar**

Extender o ampliar una cosa. Las ANN generalizan, automáticamente, debido a su propia estructura y naturaleza. Estas redes pueden ofrecer, dentro de un margen, respuestas correctas a entradas que presentan pequeñas variaciones debido a los efectos de ruido o distorsión.

- **Abstraer**

Aislar mentalmente o considerar por separado las cualidades de un objeto. Algunas ANN son capaces de abstraer la esencia de un conjunto de entradas que aparentemente no presentan aspectos comunes o relativos (p.2).

Ventajas que ofrece la red neuronal.

Las ventajas que ofrecen las redes neuronales según indica, Matich (2001), son:

Debido a su constitución y a sus fundamentos, las redes neuronales artificiales presentan un gran número de características semejantes a las del cerebro. Por ejemplo, son capaces de aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante, etc. Esto hace que ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando en múltiples áreas (p.8)

Entre las ventajas, se incluyen:

- **Aprendizaje adaptativo**

Matich (2001) afirma que la capacidad de aprendizaje adaptativo es una de las características más atractivas de redes neuronales. Esto es, aprender a llevar a cabo ciertas tareas mediante un entrenamiento con ejemplos ilustrativos.

Como las redes neuronales, pueden aprender a diferenciar patrones mediante ejemplos y entrenamientos, no es necesario elaborar

modelos a priori ni necesidad de especificar funciones de distribución de probabilidad.

Las redes neuronales son sistemas dinámicos auto adaptativos. Son adaptables debido a la capacidad de autoajuste de los elementos procesales (neuronas) que componen el sistema. Son dinámicos, pues son capaces de estar constantemente cambiando para adaptarse a las nuevas condiciones.

En el proceso de aprendizaje, los enlaces ponderados de las neuronas se ajustan de manera que se obtengan ciertos resultados específicos. Una red neuronal no necesita un algoritmo para resolver un problema, ya que ella puede generar su propia distribución de pesos en los enlaces mediante el aprendizaje. También existen redes que continúan aprendiendo a lo largo de su vida, después de completado su período de entrenamiento (p.9).

- **Autorganización**

Matich (2001) afirman que las redes neuronales emplean su capacidad de aprendizaje adaptativo para autoorganizar la información que reciben durante el aprendizaje y/o la operación. Mientras que el aprendizaje es la modificación de cada elemento procesal, auto organización consiste en la modificación de la red neuronal completa para llevar a cabo un objetivo específico.

Cuando las redes neuronales se usan para reconocer ciertas clases de patrones, ellas auto organizan la información usada. Por ejemplo, la red llamada backpropagation, creará su propia representación característica, mediante la cual puede reconocer ciertos patrones.

Esta auto organización provoca la generalización: facultad de las redes neuronales de responder, apropiadamente, cuando se les presentan datos o situaciones a la que no había sido expuesta anteriormente. El sistema puede generalizar la entrada para obtener una respuesta. Esta característica es muy importante cuando se tiene que solucionar problemas en los cuales la información de entrada no es muy clara;

además, permite que el sistema dé una solución, incluso cuando la información de entrada está especificada de forma incompleta (p.10).

- **Tolerancia a fallos**

Matich (2001) al respecto acota, “Las redes neuronales fueron los primeros métodos computacionales con la capacidad inherente de tolerancia a fallos. Comparados con los sistemas computacionales tradicionales, los cuales pierden su funcionalidad cuando sufren un pequeño error de memoria, en las redes neuronales, si se produce un fallo en un número no muy grande de neuronas y aunque el comportamiento del sistema se ve influenciado, no sufre una caída repentina”(p.11).

- **Operación en tiempo real**

Matich (2001) afirma que una de las mayores prioridades, casi en la totalidad de las áreas de aplicación, es la necesidad de realizar procesos con datos de forma muy rápida. Las redes neuronales se adaptan bien a esto debido a su implementación paralela. Para que la mayoría de las redes puedan operar en un entorno de tiempo real, la necesidad de cambio en los pesos de las conexiones o entrenamiento es mínima (p.11).

- **Fácil inserción dentro de la tecnología existente**

Matich (2001) indicó Una red individual puede ser entrenada para desarrollar una única y bien definida tarea (tareas complejas, que hagan múltiples selecciones de patrones, requerirán sistemas de redes interconectadas). Con las herramientas computacionales existentes (no del tipo PC), una red puede ser rápidamente entrenada, comprobada, verificada y trasladada a una implementación hardware de bajo coste. Por lo tanto, no se presentan dificultades para la inserción de redes neuronales en aplicaciones específicas, por ejemplo, de control, dentro de los sistemas existentes. De esta manera, las redes neuronales se pueden utilizar para mejorar sistemas en forma incremental y cada

paso puede ser evaluado antes de acometer un desarrollo más amplio (p.11).

- **La neurona artificial**

Palmer y Montaña (1999) señalaron: Las neuronas biológicas (figura 5) se caracterizan por su capacidad de comunicarse. Las dendritas y el cuerpo celular de la neurona reciben señales de entradas excitatorias e inhibitorias de las neuronas vecinas; el cuerpo celular las combina e integra y emite señales de salida. El axón transporta esas señales a los terminales axónicos, que se encargan de distribuir información a un nuevo conjunto de neuronas. Por lo general, una neurona recibe información de miles de otras neuronas y, a su vez, envía información a miles de neuronas más (p. 245).

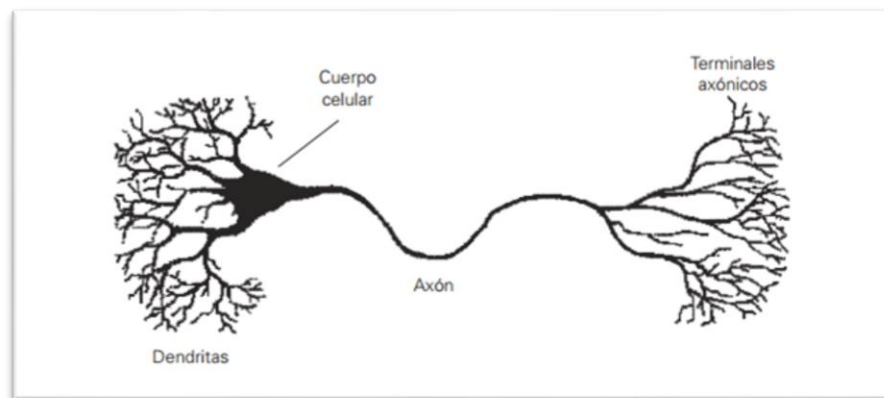


Figura 5: Estructura general de una neurona biológica.

Fuente: Matich (2001).

Palmer et al. (1999) señalaron que la neurona artificial pretende mimetizar las características más importantes de la neurona biológica. En general, recibe las señales de entrada de las neuronas vecinas ponderadas por los pesos de las conexiones. La suma de estas señales ponderadas proporciona la entrada total o neta de la neurona y, mediante la aplicación de una función matemática — denominada función de salida—, sobre la entrada neta, se calcula un valor de salida, que enviado a otras neuronas (figura 6). Tanto los valores de entrada a la neurona como la salida pueden ser señales excitatorias

(cuando el valor es positivo) o inhibitorias (cuando el valor es negativo) (p.246).

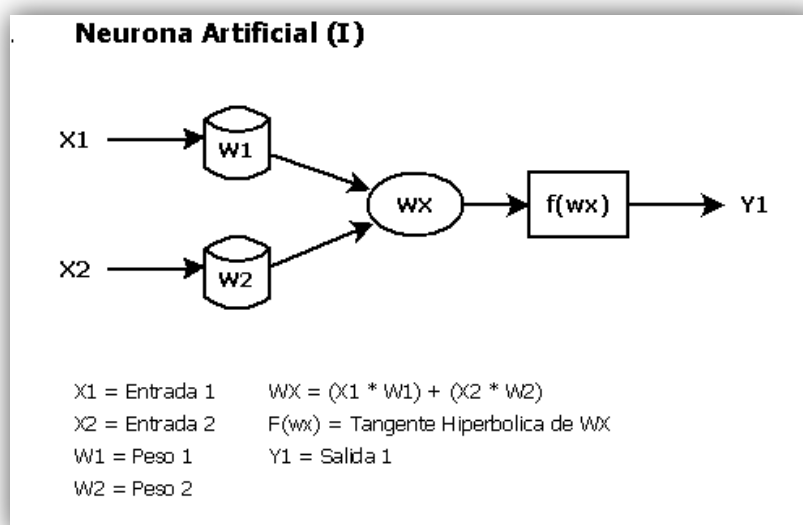


Figura 6: Funcionamiento general de una Neurona artificial.

Fuente: Matich (2001)

1.2.6.3 Árboles de decisión

El árbol de decisión es un gráfico que nos sirven como herramienta para la toma de decisiones en la empresa. Platean el problema para que todas las opciones sean analizadas, y hace posible analizar las consecuencias de adoptar una u otra decisión. También nos permite cuantificar su coste y las probabilidades de ocurrencia de cada decisión según refirió Molero (2008) pueden aplicarse, en muchas situaciones de la empresa, a la hora de la toma de decisiones, como en inversión, reinversión, políticas de créditos y financiación a corto y largo plazo (p. 24).

De las técnicas de aprendizaje, es el método más fácil de utilizar y entender. Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera, que la decisión a tomar, se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta sus hojas.

Se utilizan comúnmente cuando se necesitan detectar reglas del negocio que puedan ser fácilmente traducidas al lenguaje natural o SQL, o en la construcción de modelos de clasificación. Existen dos tipos de árboles: los de clasificación, mediante los cuales un registro es asignado a una clase en particular, reportando una probabilidad de pertenecer a esa clase, y los árboles de regresión, que permiten estimar el valor de una variable numérica objetivo.

1.2.6.4. Técnicas bayesianas

Una red bayesiana, red de Bayes, red de creencia, modelo bayesiano (de Bayes) o modelo probabilístico en un gráfico acíclico dirigido es un modelo gráfico probabilístico (un tipo de modelo estático) que representa un conjunto de variables aleatorias y sus dependencias condicionales, a través de un gráfico acíclico dirigido (DAG por sus siglas en inglés). Por ejemplo, una red bayesiana puede representar las relaciones probabilísticas entre enfermedades y síntomas. Dados los síntomas, la red puede ser usada para computar la probabilidad de la presencia de varias enfermedades.

Molero (2008) indicó que las redes bayesianas son grafos dirigidos acíclicos cuyos nodos representan variables aleatorias en el sentido de Bayes: las mismas pueden ser cantidades observables, variables latentes, parámetros desconocidos o hipótesis. Las aristas representan dependencias condicionales; los nodos que no se encuentran conectados representan variables las cuales son condicionalmente independientes de las otras. Cada nodo tiene asociado una función de probabilidad que toma como entrada un conjunto particular de valores de variables padres del nodo y devuelve la probabilidad de la variable representada por el nodo. Por ejemplo, si por padres son m variables booleanas entonces la función de probabilidad puede ser representada por una tabla de 2^m entradas, una entrada para cada una de las 2^m posibles combinaciones de los padres siendo verdadero o falso. Ideas similares pueden ser aplicadas a grafos no dirigidos, y posiblemente cíclicos; como son las llamadas redes de Markov.

Existen algoritmos eficientes que llevan a cabo la inferencia y el aprendizaje en redes bayesianas. Las redes bayesianas que modelan secuencias de variables (ej señales del habla o secuencias de proteínas) son llamadas redes bayesianas dinámicas. Las generalizaciones de las redes bayesianas que pueden representar y resolver problemas de decisión bajo incertidumbre son llamados diagramas de influencia (p. 25).

1.2.7 Procesos de la Minería de Datos

El proceso de Minería de datos consta de 06 etapas bien definidas:

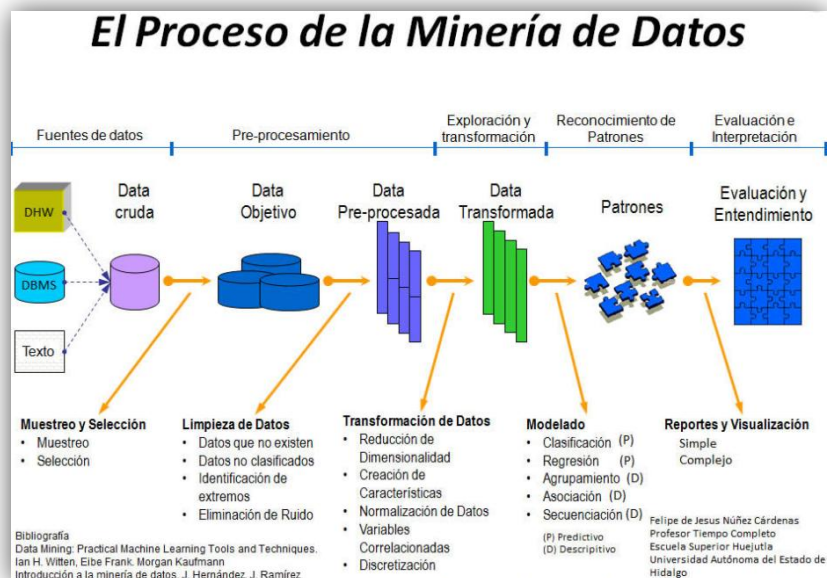


Figura 7: Proceso de minería de datos.

Fuente: KASPERU (2010).

1.2.7.1 Fuentes de datos

Las fuentes de datos según KASPERU (2010) corresponde a las fuentes originales en que se encuentran la información (bases de datos, Data Warehouse, data transaccional, archivos de textos, archivos planos, etc.) (p. 20).

Este proceso se encarga de la recolección de datos, transformados y almacenados para tenerlos en un medio uniforme.

Los datos pueden ser estructurados (Archivos Excel, Bases de datos) y no estructurados (imágenes, web mining, series de tiempo, video, voz, texto).

1.2.7.2. Pre- procesamiento – Descripción de Datos

La etapa de descripción de datos según KASPERU(2010) tiene como tareas principales, examinar las propiedades gruesas de los datos, hacer el análisis descriptivo univariado y multivariado de los datos.

El resultado de esta etapa son la visualización de datos, correlaciones estadísticas, frecuencias, formato de registros de datos, cantidad de datos (atributos, frecuencia), influencia de medición (precisión, frecuencia) y valores no disponibles.

Descripción de la forma

Histograma: Representación pictórica de la distribución de frecuencias donde el número de casos por tipo se representa en el eje vertical (se llama gráfico de barras o diagrama de bloques).

Medidas de tendencia central

Media: Valor promedio de los datos

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

La media podada: Es la media, pero sin considerar datos extremos.

La mediana: Valor que separa en dos las observaciones ordenadas de menor a mayor.

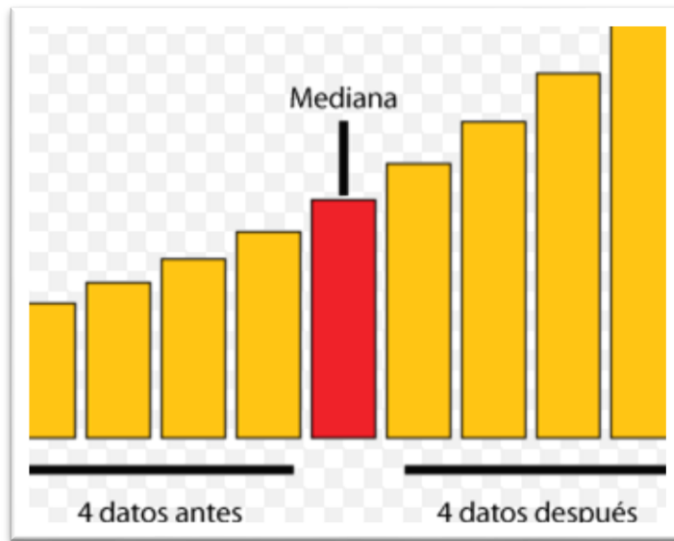


Figura 8: Medidas de Tendencia Central - La Mediana

Fuente: Extraído del artículo: Probabilidad y estadística dinámica.

La moda: Valor que más veces se repite.

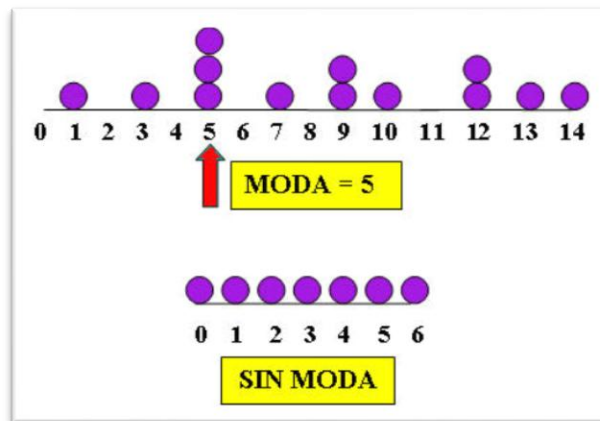


Figura 9: Medidas de Tendencia Central - La Moda

Fuente: Extraído del artículo: Probabilidad y estadística dinámica.

1.2.7.3. Pre- procesamiento – Limpieza transformación

La etapa de pre procesamiento según KASPERU (2010) corresponde al muestreo de datos, la selección de variables, limpieza de datos (datos que no existen, datos no clasificados, identificación de extremos, eliminación de ruidos, etc). (p. 20)

Identificar problemas de calidad de datos

- Datos extremos, Son aquellos que no siguen el comportamiento general de los datos. Pueden ser erróneos o correctos, pero diferentes a los demás.

Se deben identificar para eliminarlos o para sacar información de ellos (nuevos casos).

- Datos perdidos, Son los datos que no siempre están disponibles, cuando el registro no tiene un valor registrado para un atributo. Los datos desaparecidos pueden deberse a equipos que funcionaron mal, inconsistencia con otros datos registrados y así se borró, cuando los datos no fueron ingresados originalmente, no se registró la historia o el cambio de los datos.

- Datos inconsistentes

Cuando los registros se encuentran duplicados, datos inconsistentes cuando los mismos valores en el atributo, pero difieren en el valor del atributo clase.

Es un problema grave que afecta a varios métodos de aprendizaje predictivo.

1.2.7.4 Exploración y transformación

En la etapa de exploración y transformación según KASPERU (2010) se caracteriza por la creación de nuevas variables, normalización de datos, variables correlacionadas, discretización de datos, etc. (p. 10)

Los datos son transformados de forma apropiada para la extracción de información.

Existen diferentes técnicas de transformación:

- Sumarización de datos.
- Operaciones de agregación, etc.
- Aplicación de funciones a una columna (Log, sin, exp, tan, etc.)
- Aplicación de funciones entre columnas.

Técnicas de transformación de datos

Normalización min-max

$$V' = \frac{V - V_{\text{Min}}}{V_{\text{Max}} - V_{\text{Min}}}$$

Normalización Z-score

$$V' = \frac{V - m_{\text{mean}}}{\text{Std}}$$

Logaritmo natural, en base 3 o en base 10

$$V' = \text{Log}_{10}(V)$$

1.2.7.5 Reconocimiento de patrones:

Según KASPERU (2010), se construye un modelo, que será utilizado sobre los datos para predecir las clases mediante clasificación o para descubrir grupos similares mediante segmentación (p. 10).

1.2.7.6 Evaluación e interpretación:

Según KASPERU (2010) una vez aplicado el paso anterior, se buscan patrones de comportamiento en los valores de las variables del problema o relaciones de asociación entre dichas variables (p. 11).

1.2.7.7 Despliegue:

Según KASPERU (2010), el modelo debe ser validado comprobando que las conclusiones arrojadas son válidas y satisfactorias. Si el modelo final no supera esta evaluación, el proceso puede repetirse desde el principio o a partir de cualquiera de los pasos anteriores (p. 12).

1.2.8 Aplicaciones de la minería de datos

Según la prestigiosa comunidad de Mineros de Datos KDNuggets (2012), mencionaron que la aplicación de minería de datos en detección de fraudes se encuentra dentro de las diez primeros más desarrollados como vemos en la figura extraída de su página web.

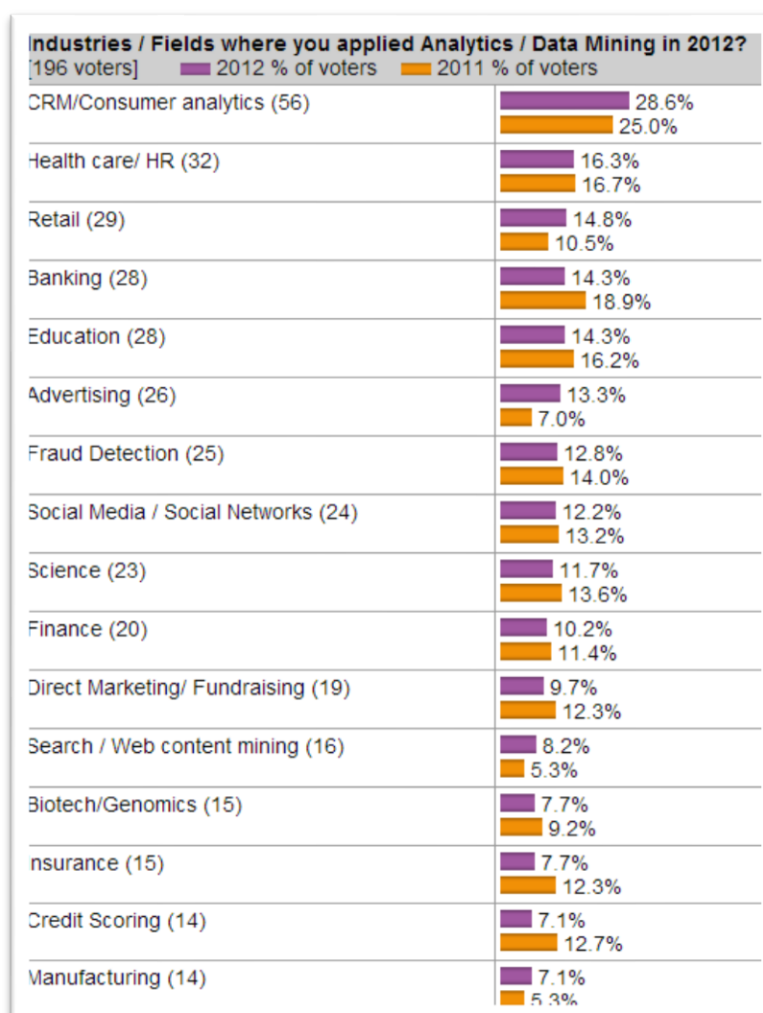


Figura 10: Encuesta en que industria se aplica DM en el 2012

Fuente: kdnuggets 2010

1.2.9 Metodologías de Minería de Datos

En el mercado actual, existen diversas técnicas y herramientas disponibles de minería de datos para el desarrollo de un proyecto. Por esta razón, diversas empresas y consultorías en el mundo han desarrollado metodologías de trabajo para guiar al usuario.

Entre las metodologías para el desarrollo de Minería de datos más importantes tenemos:

1.2.9.1. Metodología SEMMA

El acrónimo SEMMA surge de las iniciales de las palabras Sample (muestra), Explore (explorar), Modify (modificar), Model (modelar) y Assess (evaluar).

Según Camargo, Silva (2010), indicaron que SEMMA es una organización para el manejo de una herramienta funcional de SAS llamada Enterprise Manager para el manejo de tareas de minería de datos. También indicaron que SEMMA intenta hacer fácil de aplicar la exploración estadística y la visualización de técnicas, seleccionando y transformando las variables predictivas más relevantes, modelándolas para obtener resultados, y finalmente confirmar la precisión del modelo.

Las fases de la Metodología SEMMA son las que mostramos, a continuación, en la figura.

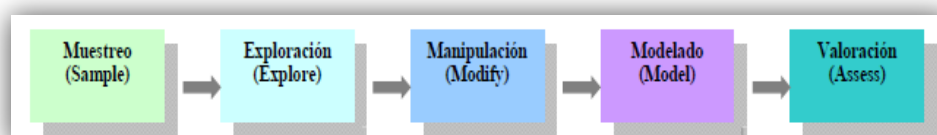


Figura 11: Fases de la Metodología SEMMA

Fuente: SAS (2014)

A Fase muestral

Este proceso se inicia con la extracción de la población muestral sobre la que se va a aplicar el análisis. El objetivo de esta fase consiste en seleccionar una muestra representativa del problema en estudio. La representatividad de la muestra es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan de ser admisibles. La forma más común de obtener una muestra es la selección al azar, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido. Este método de muestreo se denomina muestreo aleatorio simple.

Según Camargo, Silva (2010), En esta fase, se busca extraer una porción de datos lo suficientemente grande para contener información significativa, pero reducida para manipularla rápidamente. Si los patrones generales aparecen en los datos en su conjunto, estos se pueden distinguir en una muestra representativa. Si un nicho es tan pequeño que no es representable con una muestra y aun así es tan importante que influencia la imagen completa, puede ser descubierto por medio de métodos de síntesis. También se pueden crear conjuntos de datos así:

- Entrenamiento – Usado para modelos adecuados
- Validación – Usado para comprobar.
- Prueba – Usado para obtener comprobaciones honestas y para mostrar qué tan bien puede generalizar un modelo.

B Fase Exploración

Se procede a la exploración de la información disponible con el fin de simplificar en lo posible el problema para optimizar la eficiencia del modelo. Para lograr este objetivo, se propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma, se pretende determinar cuáles son las variables explicativas que van a servir como entradas al modelo.

Según Camargo Silva (2010), indicaron que en esta fase, se desea explorar los datos buscando tendencias y anomalías imprevistas para obtener una comprensión total de los mismos. Esta fase ayuda a refinar el

proceso de descubrimiento. Si visualmente no hay un resultado claro se pueden tratar los datos por medio de técnicas estadísticas como el análisis factorial, de correspondencias y agrupaciones.

Ejemplos: la minería de datos de campañas de correo directo, el agrupamiento podría revelar grupos de compradores con distintos patrones de ordenamiento, y sabiendo esto, se crea la oportunidad de generar correos personalizados o promociones.

C Fase manipulación

La tercera fase de la metodología consiste en la manipulación de los datos, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado los datos que serán introducidos en el modelo.

Según Camargo, Silva (2010), indicaron que se manipulan o modifican los datos por medio de la creación, selección y transformación de variables, para centrar el proceso de selección del modelo. Basado en los descubrimientos en la fase de exploración, puede haber la necesidad de manipular los datos para incluir información como la de agrupamiento de compradores y subgrupos significativos, o introducir nuevas variables. También puede ser necesario buscar valores extremos (bordes) y reducir el número de variables para reducir a los más significativos.

También puede ser necesario modificar datos cuando la información “minada” cambie. Debido a que la minería de datos es un proceso dinámico e iterativo, puede actualizar los métodos o los modelos cuando esté disponible nueva información.

D Fase Modelado

El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado.

Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales (tales como análisis discriminante, métodos de agrupamiento, y análisis de regresión), así como técnicas

basadas en datos tales como redes neuronales, técnicas adaptativas, lógica fuzzy (difusa), árboles de decisión, reglas de asociación y computación evolutiva.

Según Camargo, Silva (2010) precisan que en esta fase se modelan los datos permitiendo que el software busque, automáticamente, una combinación de datos que prediga con cierta certeza un resultado deseado.

Las técnicas de modelado, en minería de datos, incluyen las redes neuronales, modelos de árboles de decisión, modelos lógicos y otros modelos estadísticos (como los análisis de serie de tiempo, razonamiento basado en memoria y componentes principales). Cada uno tiene sus fortalezas, y dependiendo de la información se debe aplicar el más adecuado según las situaciones concretas para el análisis con la minería de datos. Ejemplo, las redes neuronales son muy buenas en la conexión de relaciones no lineales de gran complejidad.

E Fase Valoración

Esta fase del proceso consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos contrastados con otros métodos estadísticos o con nuevas poblaciones muestrales.

Según Camargo, Silva (2010), indicaron que califican los datos mediante la evaluación de la utilidad y fiabilidad de los resultados del proceso de minería de datos. Una forma común de evaluación de un modelo es la de aplicar el modelo a una porción aparte de resultados obtenidos durante el muestreo. Si el modelo es válido, debería funcionar para esta muestra, así como para la muestra utilizada en la construcción del modelo. De manera similar, se puede probar el modelo nuevamente con los datos conocidos. Por ejemplo, si se sabe cuáles clientes tienen altas tasas de retención y su modelo predice la retención, puede probar si el modelo selecciona estos clientes acertadamente.

1.2.8.2 Metodología CRISP-DM

CRISP-DM es acrónimo de (Cross –Industry Estándar Process for Data mining). Esta metodología inicialmente fue desarrollada por tres empresas que iniciaron sus investigaciones en el tema de la Minería de Datos: DaimlerChrysler (luego conocido como DaimlerBenz) quien siempre implementó principios y técnicas de minería de datos en sus negocios, SPSS quien provee servicios basados en Minería de Datos desde 1990, y NCR. También Kasperú (2011) indicó que luego otros líderes de la industria como son IBM, SAS y OHRA contribuyeron en el desarrollo de la metodología.

Camargo, Silva (2010), “La metodología CRISP – DM, está descrita en términos de un modelo de proceso jerárquico, que consiste en una serie de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): Fases, tareas genéricas, tareas especializadas e instancias de proceso”.

KASPERU (2011) indicó que la metodología CRISP –DM valida el proceso, dispone de modelos de referencia (plantillas), ayuda a planear y administrar proyectos, no tiene propietario.

También alienta la inter-operatividad de herramientas. Enfocado al negocio y al análisis técnico.

Gutiérrez (2007) determinó que la metodología de Data Mining CRISP-DM que está definida en términos de un modelo jerárquico de procesos, consiste de un conjunto de tareas descritas en 4 niveles de abstracción (desde lo general a lo específico): Fases, Tareas Genéricas, Tareas Especializadas e Instancias de procesos (p.25).

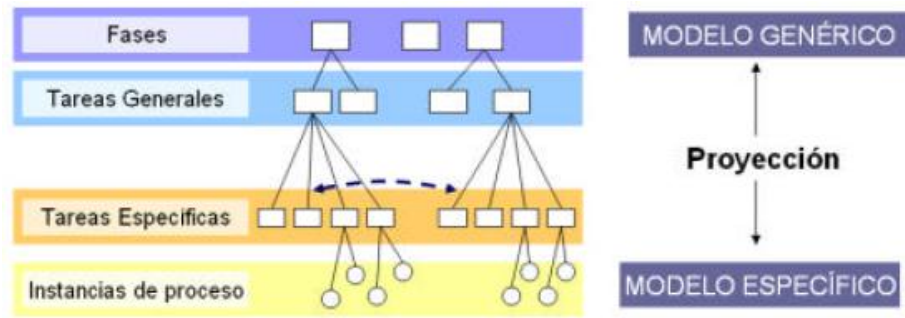


Figura 12: Esquema de Extracción de los cuatro niveles de abstracción de la metodología CRISP-DM

Fuente: Rodríguez, Álvarez, Mesa y Gonzáles (2010). Metodologías para el desarrollo de Proyectos de Data Mining

Fases

La metodología CRISP-DM está organizado en seis fases, y cada fase, a su vez, está estructurada en varias tareas generales de segundo nivel.

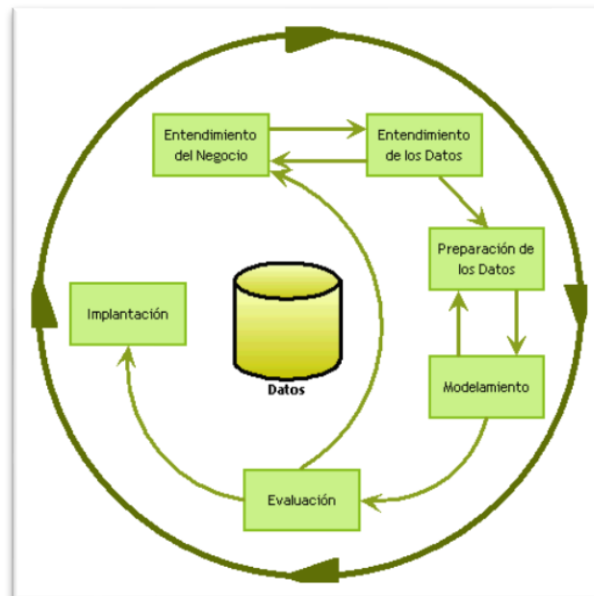


Figura 13: Metodología de Minería de Datos - Fases de la Metodología CRISP-DM

Fuente: Rodríguez, Álvarez, Mesa y Gonzáles (2010).

Las fases de la metodología CRISP-DM son 06:

1 Entendimiento de negocio: Esta fase se enfoca a la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, a la

definición de un problema de minería de datos y a un plan preliminar diseñado para alcanzar los objetivos.

2 Entendimiento de datos: Esta fase pretende coleccionar los datos iniciales, desarrollar procedimientos para entender los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

3 Preparación de datos: Esta fase cubre todas las actividades necesarias para construir el conjunto de datos que serán provistos a las herramientas de modelado desde los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

4 Modelado: En esta fase, elige las técnicas de modelado aplicadas, que sean calibradas. Típicamente, hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Esta fase se relaciona con la fase de evaluación, todo modelo debe ser evaluado para calcular su rendimiento.

5 Evaluación: En esta fase, se eligen los métodos de evaluación del modelo, cálculo del rendimiento, sensibilidad, especificidad, porcentaje de éxito y error.

6 Despliegue: En esta fase, ya se ha construido un modelo óptimo (datos + procedimientos) que tiene el mejor desempeño, bajo una perspectiva de análisis de datos. La finalidad de esta etapa es el usar el modelo óptimo para predecir comportamientos de nuevos registros de datos que parece que tienen la alta calidad de una perspectiva de análisis de datos.

Tareas Generales

Este segundo nivel es llamado genérico porque pretende ser lo más amplio posible de manera que cubra una vasta gama de situaciones de Data Mining. A su vez, las tareas deben ser lo más completas y estables posible. En este nivel, se describen las acciones que deben ser desarrolladas para situaciones específicas.

Ejemplo: Si estuviéramos en la fase de preparación de Datos, una tarea general puede ser la Limpieza de Datos.

Tareas Específicas

El tercer nivel comprende las tareas especializadas en él, se describen cómo las acciones de las tareas genéricas se deben efectuar en situaciones específicas.

Ejemplo: Si en el segundo nivel puede haber una tarea general llamada limpiar datos; en cambio, en el tercer nivel se describe como esta tarea se lleva a cabo en diferentes situaciones, tales como limpiar valores numéricos versus limpiar valores categóricos o si el tipo de problema es clustering o modelamiento predictivo

Instancias de Procesos

El cuarto nivel, la instancia de los procesos, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de Data Mining específico.

Una instancia de proceso es organizada de acuerdo a las tareas definidas en los niveles superiores, pero grafica que está ocurriendo realmente en un compromiso en particular, más que lo que pasa en general.

La descripción de fases y tareas como pasos discretos efectuados en orden específico representa una secuencia idealizada de eventos. En la práctica, muchas de las tareas pueden llevarse a cabo en distinto orden y a menudo será necesario volver atrás repetidamente a las tareas previas (backtracking) y repetir ciertas acciones.

1.2.9 Comparación de metodologías

Según el estudio realizado por Rodríguez, Álvarez, Mesa y González (2010), investigan sobre las metodologías para desarrollar y gestionar proyectos de minería de datos indicaron que:

Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de Data Mining en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Data Mining en un proceso iterativo e interactivo.

La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto. Esta diferencia se establece ya desde la primera fase del proyecto de Data Mining donde la metodología SEMMA comienza realizando un muestreo de datos, mientras que la metodología CRISP-DM comienza realizando un análisis del problema empresarial para su transformación en un problema técnico. Desde ese punto de vista más global se puede considerar que la metodología CRISP-DM está más cercana al concepto real de proyecto.

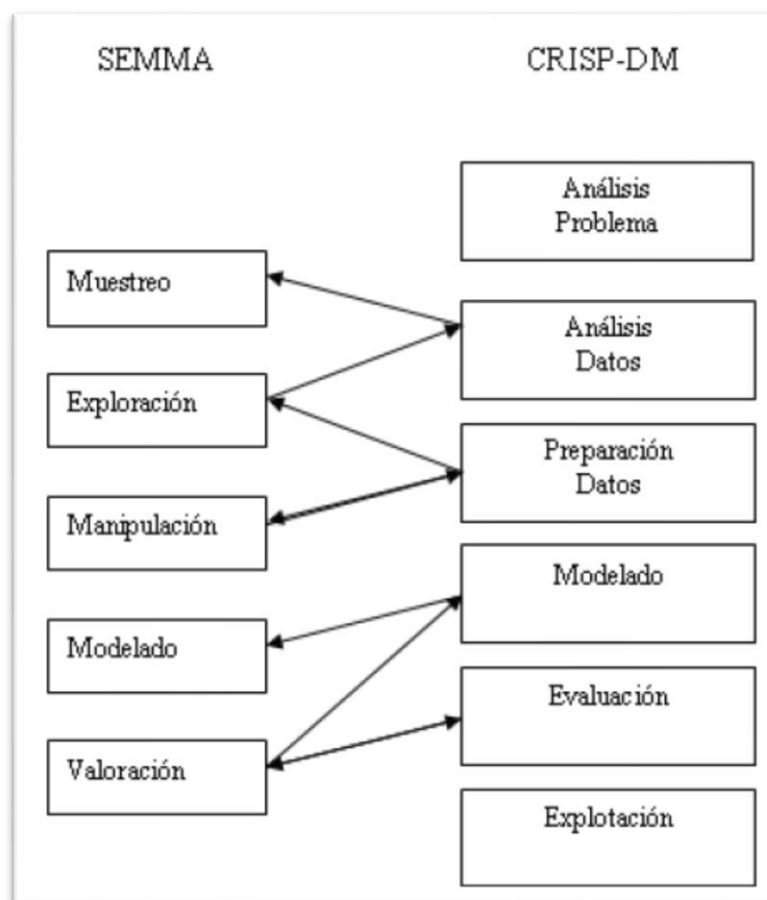


Figura 14: Comparativa de las interrelaciones entre las fases de las metodologías SEMMA y CRISP-DM

Fuente: Rodríguez, Álvarez, Mesa y Gonzáles (2010). Metodologías para el desarrollo de Proyectos de Data Mining

1.3 Definición de términos básicos

- Acometida: Instalación por la que se deriva hacia un edificio u otro lugar parte del fluido que circula por una conducción principal
- Comercialización: Dar a un producto, condiciones y vías de distribución para su venta.
- Contrato de servicio: Documento que firma la empresa y el usuario final.
- Componentes del sistema de distribución eléctrica: Compuesto por los generadores o alternadores, alimentadores, transformadores, líneas y cables, capacitadores o condensadores y equipos de protección.
- Generador eléctrico: Es todo dispositivo capaz de mantener una diferencia de potencial eléctrico entre dos de sus puntos, llamados polos, terminales o bornes. Los generales eléctricos son máquinas destinadas a transformar la energía mecánica en eléctrica.
- Líneas aéreas: Son las líneas de distribución eléctrica típicas líneas en las que los cables van colgados de postes ya sean de madera o de metal, se suelen usar para reducir costes, ya que nos ahorramos el coste del aislante al ir los cables desnudos, siendo el aislante el propio aire que separa las fases, y los costes que supondría tener que hacer las canalizaciones en el suelo, otra ventaja es que es más fácil ver donde se ha roto la línea.
- Líneas enterradas: Son líneas de distribución eléctrica donde los cables están recubiertos por sus correspondientes aislantes, y van a través de canalizaciones que pueden ser tubos de plástico o metálicos, canales de cemento, zanjas excavadas en la tierra, colgados de paredes en túneles.
- Modelo: Representación de la realidad.
- Nodo: Punto de inflexión.

- **Potencia eléctrica:** La potencia eléctrica es la relación de paso de energía de un flujo por unidad de tiempo; es decir, la cantidad de energía entregada o absorbida por un elemento en un tiempo determinado. La unidad en el Sistema Internacional de Unidades es el vatio (watt).
- **Sistema eléctrico:** Es una serie de elementos o componentes eléctricos o electrónicos, tales como resistencias, inductancias, condensadores, fuentes, y/o dispositivos electrónicos semiconductores, conectados eléctricamente entre sí con el propósito de generar, transportar o modificar señales electrónicas o eléctricas.
- **Sistema de Distribución Eléctrica:** Es parte del sistema de suministro eléctrico cuya función es el suministro de energía desde la subestación de distribución hasta los usuarios finales (medidor del cliente).
- **Subestación:** Es usada para la transformación de la tensión de red o de nuestro generador a una tensión adecuada a nuestras necesidades.
- **Subestación de intemperie:** Son las encargadas de regular y gestionar el transporte de la energía eléctrica, su aislante es el aire o espacio que hay entre los elementos.
- **Subestaciones blindadas:** La más usual es GIS, Gas Insulated Switchgear. En ellas el fluido que trabaja como aislante es el gas SF6, hexafluoruro de azufre. Usado en la mayoría de interruptores de subestaciones eléctricas convencionales por sus adecuadas características para la eliminación del arco eléctrico.
- **Transformador:** El transformador de tensión es el principal elemento de la subestación, es el encargado de convertir el valor de la tensión del generador en el valor de la tensión de la red donde volcamos la energía producida, por lo que es un punto crítico al ser por donde sale toda la energía eléctrica.

- Transmisión: Conjunto de mecanismos que comunican el movimiento de un cuerpo a otro, alterando generalmente su velocidad, su sentido o su forma.

CAPÍTULO II

METODOLOGÍA

2.1 Aspectos metodológicos de la investigación

- **Variables independientes**

Datos de Consumidores de energía eléctrica, lo cuales se encuentran en la tabla BaseConsumo .Entregada por la empresa eléctrica.

Tabla 1: Variables independientes de Consumidores de energía eléctrica.

Variable	Tipo de Dato	Descripción
NumeroCuenta	Texto	Indica el número de cuenta del cliente
NumeroServicio	Texto	Indica el número de servicio de la cuenta. Es igual al número de cuenta.
Nombre	Texto	Nombre del cliente
Direccion	Texto	Dirección del cliente
CodigoDistrito	Texto	Código del distrito de residencia del cliente
Distrito	Texto	Nombre del distrito de residencia del cliente
Sector	Texto	Sector en el que se encuentra ubicada la casa del cliente
Zona	Texto	Zona a la que pertenece el sector
Correlat	Texto	-
SucursalComercial	Texto	-
CodigoActividadComercial	Texto	Código de la actividad comercial a la que se dedica el cliente.
ActividadComercial	Texto	Nombre de la actividad comercial a la que se dedica el cliente
Tarifa	Texto	-

TipoRedElectrica	Texto	Tipo de red eléctrica a la que pertenece el cliente.
Potencia	Texto	-
TipoCliente	Texto	Tipo de cliente
EstadoServicio	Texto	-
EstadoCliente	Texto	-
Fecha_Ult_Deteccion	Date	Fecha en que ha sido detectado como hurtador por última vez en todo su historial.
Codigolrregularidad_Ult_Deteccion	Texto	código de irregularidad de la última detección de hurto
Irregularidad_Ult_Deteccion	Texto	Nombre de la Irregularidad de la última detección de hurto
NumeroNotificacion_Ult_Deteccion	Texto	Numero de notificación de la última detección de hurto
Año_NumeroExpediente	Texto	Año del número del expediente.
Medidor	Texto	Medidor que tiene instalado el cliente.
Marca	Texto	Marca del medidor
Modelo	Texto	Modelo del medidor
Fase	Texto	-
Factor	Texto	-
[Estado Medidor]	Texto	Estado en que se encuentra el medidor.
FecInstalacion	Date	Fecha de instalación del medidor.
[Sello Medidor]	Texto	-
[Sello Bornera]	Texto	-
[Sello Contraste]	Texto	-
[SET]	Texto	-
Alimentador	Texto	-
SED	Texto	-
Llave	Texto	-
FecLectura_i	Date	Fecha de lectura del mes "i". Se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Correspondiente al consumo del cliente
[Dias Fact_i]	Numero	Días facturados del mes "i". Se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Correspondiente al consumo del cliente
[Clave Lect_i]	Texto	Clave de lectura del mes "i". Se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Correspondiente al consumo del cliente
[Clave Fact_i]	Texto	Clave de facturación del cliente para el mes número "i"
Consumo_i	Numero	Consumo del mes "i" se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Medido en Kw

Elaboración: La autora

Datos de Consumidores Fraudulentos. Igualmente corresponde a información entregada por la empresa eléctrica.

Tabla: Valorizaciones.

Tabla 2: Variables Independientes de Consumidores fraudulentos

Variable	Tipo de Dato	Descripción
NumeroCuenta	Texto	Indica el número de cuenta del cliente
NumeroServicio	Texto	Indica el número de servicio de la cuenta. Es igual al número de cuenta.
NumeroInspeccion	Texto	Indica el número de la inspección realizada
AñoExpediente	Texto	Indica el año del expediente.
NumeroExpediente	Texto	Indica el número del expediente.
FechaCreacion	Date	Señala la fecha de creación de la valorización
FechaInicioRecupero	Date	Indica la fecha de inicio del recupero. La fecha en que comenzó a hurtar
FechaFinRecupero	Date	Indica la fecha de fin de recupero. La fecha en que ya se controló el problema y ya no está hurtando.
TotalEnergia	Numero	Señala el total de energía hurtada.
TotalRecupero	Numero	Señala el total recuperado.
TotalIntereses	Numero	-
TotalMoras	Numero	-

Elaboración: La autora

- **Variables Dependientes**

Se obtiene transformando en base a la fecha de inicio y fin de recupero de la tabla valorizaciones.

Tabla 3: Variable dependiente

Variables	Tipo de Datos	Descripción
Hurtold	Texto	Esta variable corresponde al resultado de la predicción, tiene dos valores SI (hurto) No (No hurto).

Elaboración: La autora

2.2 Tipos de Metodología

La investigación que se desarrolló fue aplicada, explicativa, longitudinal y cuantitativa.

Aplicada

La investigación es aplicada porque se propone un modelo matemático automatizado que disminuya los fraudes de energía eléctrica en clientes residenciales y contribuya con esta solución con el área de recupero de energía que es un problema social que afecta a las empresas y calidad de vida de los consumidores formales nos apoyamos en lo expuesto por el autor Ander-Egg (1995) donde indicó que “La investigación prácticas o aplicadas tiene como propósito estudiar aspectos de la realidad, comprobación de hipótesis y busca la solución de determinados problemas sociales, está íntimamente ligada a la investigación básica pues depende de descubrimientos, avances y se enriquece con ellos”. (p. 16)

Explicativa

El proyecto propuesto es explicativo porque en base al análisis de datos se logra obtener las variables explicativas que permita crear un modelo con alto grado de asertividad basado en redes neuronales que aprenda los comportamientos de consumidores fraudulentos y no fraudulentos de energía eléctrica dando una probabilidad que determine la sospecha, nos apoyamos en el sustento del autor Sabino (1996) quien indicó que “Los

trabajos son explicativos porque muestra preocupación en los orígenes y causas de un determinado fenómeno, cuyo objetivo es conocer porque suceden ciertos hechos, analiza las relaciones causales existentes y las condiciones que ellos producen. En este tipo de investigación, se explica la razón y el porqué de las cosas”. (p. 63)

Longitudinal

El modelo corresponde a una investigación longitudinal debido a que utiliza la historia de los consumidores en un intervalo de tiempo para conocer y entender su comportamiento. Esta definición la apoyamos en los autores Rodríguez y Llorca (2004) quienes postulan que un estudio es considerado longitudinal cuando se basa en la experiencia de la población a lo largo del tiempo.

Cuantitativa

El modelo propuesto corresponde a una investigación cuantitativa debido a que los datos utilizados fueron numéricos y en la etapa de comprensión de datos utilizamos técnicas estadísticas en el análisis de datos, así como análisis multivariado de las variables para poder identificar que variables apoyan y explican mejor el comportamiento de los consumidores y poder elegir que datos se utilizarán en el desarrollo del modelo, no apoyamos en lo sugeridos por los autores García y Martínez (2000) quienes indicaron “Los métodos de investigación cuantitativa utiliza registros estructurados de observación y técnicas estadísticas de análisis de datos, se centra en el estudio de las relaciones entre variables cuantificadas”. (p. 215).

2.3 Recursos

2.3.1. Recursos humanos

a. Definición de recursos

Líder del proyecto: Responsable de detectar las necesidades de los usuarios y gestionar los recursos económicos, materiales y humanos, para obtener los resultados esperados en los plazos previstos y con la calidad necesaria:

Responsabilidades:

- Identificando, haciendo seguimiento y resolviendo los asuntos del proyecto.
- Identificando, gestionando y mitigando el efecto de los riesgos.
- Asegurando que el resultado-producto del proyecto tenga una calidad adecuada.
- Administrando pro activamente el alcance para asegurar que únicamente lo acordado sea entregado, a menos que los cambios hayan sido aprobados mediante un proceso de manejo de cambio de alcance
- Definiendo y recopilando información estadística-métrica para dar sentido práctico a como el proyecto está progresando y que los productos entregados sean aceptables.
- Gestionar el plan de trabajo a fin de asegurar que las tareas sean asignadas y terminadas a tiempo y dentro del presupuesto.

Usuario especialista de recupero de energía eléctrica: posee dominio del sistema de distribución de energía eléctrica y de los procesos y procedimientos involucrados en el recupero de energía eléctrica, propone métodos de solución.

Analista de Minería de Datos: La función del analista de minería de datos consiste en consolidar, desarrollar e interpretar los resultados de las actividades de minería de datos.

Responsabilidad:

- Desarrollo de la herramienta predictiva de Minería de Datos.
- Identificar patrones de comportamiento de los datos que lleven a predecir el fraude.
- Análisis e interpretación de resultados de la aplicación de Minería de Datos.

b. Asignación de recursos

Tabla 4: Recursos del Proyecto

Responsabilidad	Cantidad	Recurso
Empresa Eléctrica	01	líder del Proyecto
Empresa Eléctrica	01	Un especialista del área de recupero de energía eléctrica
Consultora	01	líder del Proyecto
Consultora	01	Un analista de Minería de Datos
Consultora	01	Un Analista Soporte Post Implementación.

Elaboración: La autora

2.3.2. Recursos herramientas.

Se trabajará con las herramientas que cuenta la empresa. Por lo tanto se asumo costo “cero” en dichos recursos.

Tabla 5: Hardware y Software del Proyecto

Cantidad	Hardware	Software
01	Pc escritorio ó Laptop	Sistema Operativo: Window Xp professional ó superior
01		Licencia Sql server 2008. Database Engire. Analysis services
01		Licencia Microsoft Office 2007 ó Superior
01		Addins para Microsoft Office 2007 para minería de datos (gratuito)

Elaboración: La autora

2.4 Evaluación de la Viabilidad Económica

Inversión Inicial:

Tabla 6: Inversión Inicial del Proyecto

Cantidad	Personal	Consto/Hora	Total Horas	Costo
01	Líder del Proyecto	50	300	S/15,000.00
01	Analista de Minería de Datos	30	300	S/9,000.00
01	Soporte Post	30	100	<u>S/3,000.00</u>
Total				S/27,000.00

Elaboración: La autora

Monto aproximado a recuperar con el proyecto

Tabla 7: Monto a recuperar con el Proyecto

VP %	Cantidad VP	Costo	Consumo detectado KW	Costo Total Soles por Recuperar
55%	10399	0.38	948,472.48	S/360,419.54

Elaboración: La autora

Viabilidad Económica Proyectada a los próximos 5 Periodos

Margen de Contribución: 46%

Inversiones Adicionales 27,000.00

Tabla 8: Ingresos Adicionales del Proyecto

Ingresos Adicionales	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Margen de Contribución Adicional		S/165,792.99	S/165,792.99	S/165,792.99	S/165,792.99	S/165,792.99
Ahorros						
Total	0	S/165,792.99	S/165,792.99	S/165,792.99	S/165,792.99	S/165,792.99

Elaboración: La autora

Tabla 9: Egresos Adicionales del Proyecto

Egresos Adicionales	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Inversiones Adicionales	(S/27,000)					
Gastos		S/5,400.	S/5,400	S/5,400.00	S/5,400.00	S/5,400.00
Total	(S/27,000)	S/5,400	S/5,400	S/5,400	S/5,400	S/5,400

Elaboración: La autora

Tabla 10: Flujo Neto del Proyecto

	Año 0	Año 1	Año 2	Año 3	Año 4	Año 5
Flujo Neto	(S/27,000)	S/160,392.99	S/160,392.99	S/160,392.99	S/160,392.99	S/160,392.99

TMAR	10%
VNA	581016

Elaboración: La autora

Evaluación de Inversiones

Costo Promedio del Capital: (C.P.I.=i)

Tabla 11: Fuente de Fondos

Fuentes de Fondo	Importe	Porcentaje %	Costo	Costo Promedio
1.4 Préstamo	0	0%	0%	0%
1.5 Aporte Propio	S/27000	100%	30%	30%
	S/27000	100%		30%

Elaboración: La autora

Es decisión de los socios: 30%

Indicadores Financieros de Rentabilidad

Valor Actual Neto

Fc= Flujo de caja

Io= Inversión inicial del proyecto

n= Vida útil del proyecto

t= n años

i=Costo promedio del capital

$$VAN = \frac{Fc1}{(1+i)^1} + \frac{Fc2}{(1+i)^2} + \frac{Fc3}{(1+i)^3} + \dots + \frac{Fcn}{(1+i)^n}$$

Tabla 12: Valor Actual Neto

VAN	S/. 160,392.99	S/. 160,392.99	S/. 160,392.99	S/. 160,392.99	S/. 160,392.99	-S/ 27,000.00
	1.30	1.69	2.197	2.86	3.71	
	S/. 123,379.2232	S/. 94,907.09479	S/. 73,005.45753	S/. 56,158.04425	S/. 43,198.49558	
VAN	S/. 363,648.32					

Elaboración: La autora

Cálculo del TIR

Tabla 13: Calculo del TIR i2

	S/. 160,392.99	S/. 160,392.99	S/. 160,392.99	S/. 160,392.99	S/. 160,392.99	-S/ 27,000.00
	6.94	48.16	334.26	2319.73	16098.94	
	S/. 23,111.38	S/. 3,330.17	S/. 479.85	S/. 69.14	S/. 9.96	
VAN	S/.0.51	i2=5.92				

Elaboración: La autora

Tabla 14: Calculo del TIR i1

	S/ 160,392.99	S/ 160,392.99	S/ 160,392.99	S/ 160,392.99	S/ 160,392.99	-S/ 27,000.00
	6.95	48.30	335.70	2333.13	16215.26	
	S/ 23,078.13	S/ 3,320.59	S/ 477.78	S/ 68.75	S/ 9.89	
VAN	-S/.44.86	I1=5.93				

Elaboración: La autora

$$TIR = i1 + (i2 - i1)X \left(\frac{VAN1}{VAN2 + VAN1} \right)$$

TIR=587%

2.5 Cronograma de Actividades

El Plan de Trabajo establecido por un periodo de 82 días proyectado Inicial:

EL inicio real se movió 10 días.

Task Mod	Task Name	Duration	% Complete	% Work Complet	Actual Work	Start	Finish
	1.0- Modelo de Detección de Pérdidas No Técnicas de Energía Eléctrica en Clientes Residenciales Aplicando Minería de Datos	82 days	47%	67%	480 hrs	Sat 01/03/14	Sat 14/06/14
	+ 1.1- Planificación del Proyecto	1 day	100%	100%	0 hrs	Sat 01/03/14	Sat 01/03/14
	+ 1.2- Entendimiento de Negocio	4 days	100%	100%	40 hrs	Sun 02/03/14	Wed 05/03/14
	+ 1.3- Compresión de Datos	8 days	100%	100%	120 hrs	Thu 06/03/14	Mon 17/03/14
	+ 1.4 - Preparación de Datos	44 days	29%	59%	160 hrs	Tue 18/03/14	Fri 16/05/14
	+ 1.5 - Modelado	8 days	57%	57%	64 hrs	Sat 17/05/14	Sat 24/05/14
	+ 1.6 - Evaluación de Modelo	8 days	43%	43%	48 hrs	Sun 25/05/14	Sun 01/06/14
	+ 1.7 - Despliegue	3 days	75%	75%	48 hrs	Tue 03/06/14	Thu 05/06/14

Figura 15: Cronograma de Actividades General

Elaboración: La autora

	82 days	47%	67%	480 hrs	Sat 01/03/14	Sat 14/06/14
□ 1.0- Modelo de Detección de Pérdidas No Técnicas de Energía Eléctrica en Clientes Residenciales Aplicando Minería de Datos						
- 1.1- Planificación del Proyecto	1 day	100%	100%	0 hrs	Sat 01/03/14	Sat 01/03/14
Diseño del Plan de Trabajo aplicando la metodología CRISP_DM	0 days	100%	100%	0 hrs	Sat 01/03/14	Sat 01/03/14
Resultado - Plan de Trabajo	0 days	100%	100%	0 hrs	Sat 01/03/14	Sat 01/03/14
- 1.2- Entendimiento de Negocio	4 days	100%	100%	40 hrs	Sun 02/03/14	Wed 05/03/14
Entendimiento del negocio	1 day	100%	100%	8 hrs	Sun 02/03/14	Sun 02/03/14
Entendimiento de objetivos del problema	1 day	100%	100%	8 hrs	Mon 03/03/14	Mon 03/03/14
Entendimiento de Estructura de datos iniciales	1 day	100%	100%	8 hrs	Tue 04/03/14	Tue 04/03/14
Preparar estructuras según requerimiento	1 day	100%	100%	8 hrs	Wed 05/03/14	Wed 05/03/14
Resultado - Entendimiento de Negocio	1 day	100%	100%	8 hrs	Wed 05/03/14	Wed 05/03/14
- 1.3- Compresión de Datos	8 days	100%	100%	120 hrs	Thu 06/03/14	Mon 17/03/14
Preparar datos inicial según requerimiento	1 day	100%	100%	8 hrs	Thu 06/03/14	Thu 06/03/14
Descripción de Datos	1 day	100%	100%	16 hrs	Fri 07/03/14	Fri 07/03/14
Análisis univariado	2 days	100%	100%	32 hrs	Sun 09/03/14	Tue 11/03/14
Problemas de calidad de Datos	3 days	100%	100%	48 hrs	Wed 12/03/14	Fri 14/03/14
Entregable - Documento de Comprensión de Datos	1 day	100%	100%	16 hrs	Sat 15/03/14	Mon 17/03/14
- 1.4 - Preparación de Datos	44 days	29%	59%	160 hrs	Tue 18/03/14	Fri 16/05/14
Selección de variables	4 days	100%	100%	64 hrs	Tue 18/03/14	Fri 21/03/14
Importación de Datos	4 days	75%	0%	0 hrs	Sat 22/03/14	Wed 26/03/14
Creación de Variables	4 days	0%	0%	0 hrs	Thu 27/03/14	Tue 01/04/14
Limpieza de Datos	20 days	0%	0%	0 hrs	Wed 02/04/14	Tue 29/04/14

Figura 16: Cronograma de Actividades-Parte 1

Elaboración: La autora

1.4 - Preparación de Datos	44 days	29%	59%	160 hrs	Tue 18/03/14	Fri 16/05/14
Selección de variables	4 days	100%	100%	64 hrs	Tue 18/03/14	Fri 21/03/14
Importación de Datos	4 days	75%	0%	0 hrs	Sat 22/03/14	Wed 26/03/14
Creación de Variables	4 days	0%	0%	0 hrs	Thu 27/03/14	Tue 01/04/14
Limpieza de Datos	20 days	0%	0%	0 hrs	Wed 02/04/14	Tue 29/04/14
Extracción de patrones	12 days	42%	42%	80 hrs	Wed 30/04/14	Thu 15/05/14
Entregable - Documento de Preparación de Datos	1 day	100%	100%	16 hrs	Fri 16/05/14	Fri 16/05/14
1.5 - Modelado	8 days	57%	57%	64 hrs	Sat 17/05/14	Sat 24/05/14
Determinar datos de entrenamiento de Modelo	1 day	0%	0%	0 hrs	Sat 17/05/14	Sat 17/05/14
Determinar datos de Testeo de Datos	2 days	0%	0%	0 hrs	Sun 18/05/14	Mon 19/05/14
Creación de Modelo	2 days	100%	100%	32 hrs	Tue 20/05/14	Wed 21/05/14
Evaluación de Modelo	2 days	100%	100%	32 hrs	Thu 22/05/14	Fri 23/05/14
Entregable - Documento de Creación de Modelo	0 days	100%	100%	0 hrs	Sat 24/05/14	Sat 24/05/14
1.6 - Evaluación de Modelo	8 days	43%	43%	48 hrs	Sun 25/05/14	Sun 01/06/14
Determinación de parámetros optimos	2 days	0%	0%	0 hrs	Sun 25/05/14	Mon 26/05/14
Evaluación de resultados del modelo	3 days	67%	67%	32 hrs	Tue 27/05/14	Thu 29/05/14
Determinar Sospechosos	2 days	50%	50%	16 hrs	Fri 30/05/14	Sat 31/05/14
Entregable - Documento de Evaluación de Modelo	0 days	100%	100%	0 hrs	Sun 01/06/14	Sun 01/06/14
1.7 - Despliegue	3 days	75%	75%	48 hrs	Tue 03/06/14	Thu 05/06/14
Generación de resultado	2 days	50%	50%	16 hrs	Mon 02/06/14	Tue 03/06/14
Calculo de Sospechosos	1 day	100%	100%	16 hrs	Wed 04/06/14	Wed 04/06/14
Entregable: Documento de Despliegue del Modelo	1 day?	100%	100%	16 hrs	Thu 05/06/14	Thu 05/06/14

Figura 17: Cronograma de Actividades-Parte 2

Elaboración: La autora

CAPÍTULO III

DESARROLLO DEL PROYECTO

3.1. Esquema solución propuesta para el desarrollo del modelo predictivo

La solución propuesta, en la tesis, es desarrollar un modelo para la detección de fraudes de energía eléctrica que aprenda los diferentes comportamientos de clientes hurtadores y no hurtadores basándose en su historia de consumos y apoye en la detección de fraudes de energía eléctrica en clientes residenciales aplicando Minería de Datos.

Para crear el modelo nos basamos en la información proporcionada por una empresa distribuidora de energía eléctrica con datos de clientes residenciales de Lima Metropolitana, la empresa cuenta con la plataforma de SQL Server Enterprise 2008R2, la cual tiene como motor de base de datos al SQL Server Database Engine para crear tablas , vistas, procedimientos, funciones y almacén de datos y al SQL Server Análisis Services que contiene las características y herramientas necesarias para crear soluciones de minería de datos, Este último sirve para para crear, administrar, examinar modelos de minería de datos y crear predicciones a partir de dichos modelos y utiliza el lenguaje DMX (Extensiones de minería de datos),

que sirve para administrar modelos de minería de datos y crear consultas predictivas.

Como la empresa ya cuenta con estas herramientas se utilizó como herramienta interactiva para el usuario y fácil entendimiento para la creación, consulta y evaluación del modelo utilizar el complemento de Minería de Datos para Microsoft Excel que es gratuito y trabaja con una conexión del Análisis Services para el desarrollo de Minería de Datos.

Esquema Solución para crear el modelo predictivo

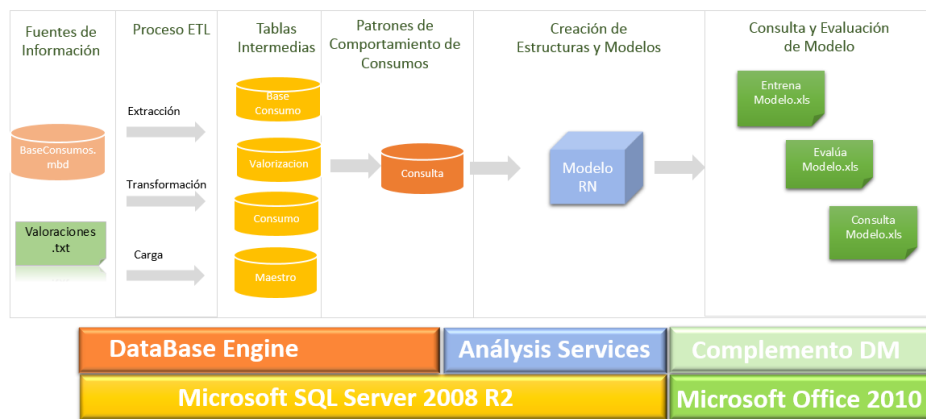


Figura 18: Esquema solución propuesta para crear el modelo predictivo de detección de fraudes de energía eléctrica en clientes residenciales

Elaboración: La autora

El esquema solución propuesta para desarrollar el modelo predictivo de detección de fraudes de energía eléctrica en clientes residenciales para su creación y desarrollo se utilizó la Metodología CRISP – DM que según investigación es la más apropiada ya que se inicia su proceso entendiendo los objetivos de la empresa.

El proceso se inicia con la importación de las fuentes de datos, la empresa brindó un archivo de Microsoft Access 2003 de nombre Base Consumos.mdb con los consumos propios de los clientes y otra archivo de Texto de nombre valorizaciones.txt.

Luego de realizar la fase de entendimiento de datos donde se evalúan los datos proporcionados, viendo su consistencia y se seleccionan las variables predictivas que apoyen para el objetivos deseado, luego en la etapa de preparación se define el tratamiento, transformación y limpieza de datos a realizar, esto se a través de un proceso ETL, basado en lenguaje SQL que lleva información hacia unas tablas intermedias dbo.Consumos (contiene la información consistente) y dbo. Maestro (contiene la información de consumos promedio diario y en cada mes indica si hubo un hurto o no); esta última, se extrae casos o patrones de comportamientos y de donde se extraerá los datos de entrenamiento, testeo de modelos y con la herramienta de Análisis Services se creará el modelo, evaluarlo y consultarlo.

1. **Fuentes de información:** Corresponde a la información original brindada por la empresa, muestras de información transaccional.
2. **Proceso ETL:** Son los procedimientos de Extracción, Transformación y Carga que se realizarán para poder preparar la información para crear los modelos, estos proceso ETL se crearán con procedimientos almacenados.
3. **Tablas intermedias:** Es el ambiente designado para tablas intermedias que tendrán el resultado de la preparación de datos.
4. **Patrones de comportamiento:** En esta parte se tiene en la tabla dbo.Consulta los casos o patrones de comportamiento que se utilizarán para entrenar, testear y consultar los modelos.
5. **Área de análisis:** Son los reportes Excel extraídos con vistas que consumen de la tabla dbo.Consulta para entrena, testear y consultar los modelos.
6. **Minería de datos:** Es el ambiente de Análisis services donde se crearan los modelos de minería de datos se utilizará el complemento

de minería de datos y se crearán las estructuras, modelos con lenguaje DMX.

3.2. Metodología CRISP – DM para la creación y desarrollo del modelo predictivo.

Como se mencionó en las bases teóricas se utilizará la Metodología CRISP-DM por ser la que mejor se adapta al proyecto por estar orientada a los objetivos en todas sus fases de desarrollo: Entendimiento de Negocio, Entendimiento de Datos, Preparación de Datos, Modelado, Evaluación y Despliegue.

Fase de minería de datos

Fase 1: Entendimiento del Negocio

Esta fase se enfoca a la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, a la definición de un problema de minería de datos y a un plan preliminar diseñado para alcanzar los objetivos.

Las empresas de Distribución Eléctricas compran energía a las empresas generadoras y la distribuyen a hogares y a empresas que requieran sus servicios. Este proceso de distribución de energía, se pueden presentar dos problemas vinculados con la pérdida.

En general, las pérdidas de energía pueden clasificarse en dos grupos:

- **Pérdidas técnicas:** Son las que se producen a causa del hecho físico que constituye la circulación de corriente eléctrica y la presencia de tensión en las redes. Pérdidas por condiciones propias de las instalaciones. Principalmente por efecto Joule.
- **Pérdidas no técnicas:** Están constituidas por :
 - La energía efectivamente suministrada pero no-medida, Energía no registrada comercialmente como tal (fraude, robo o hurto de energía, errores de facturación, errores de lectura de mediciones, etc.).
 - Consumo y no medida.

- Medida por no facturada.

Según consultores asociados (2006) se pueden clasificar según el origen puede en:

1. **Por robo o hurto:** Comprende a la energía que es apropiada ilegalmente de las redes por usuarios que no tienen sistemas de medición (conexiones clandestinas o “colgados”).
2. **Por fraude:** Corresponde a aquellos usuarios que manipulan los equipos de medición para que registren consumos inferiores a los reales.
3. **Por administración:** Corresponde a energía no registrada por la gestión administrativa de la empresa (errores de medición, errores en los procesos administrativos, falta de registro adecuada, obsolescencia de medidores, errores en los registros de censos de instalaciones de alumbrado público).

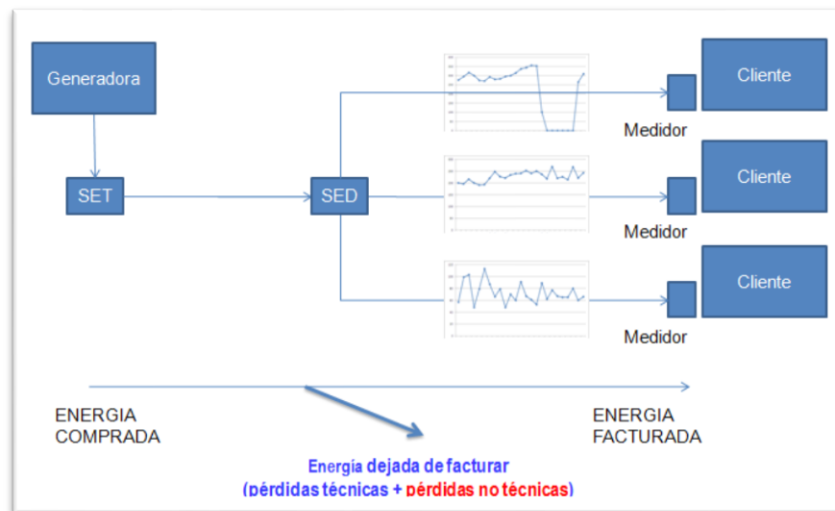


Figura 19: Pérdidas Técnicas y No Técnicas

Fuente: KASPERU (2010).

Dado que las pérdidas técnicas se pueden calcular con anticipación, no existe dificultad en su cálculo, ni en la predicción del valor monetario de la pérdida, pero en el caso de las fraudes el cálculo se dificulta, dado que corresponde por un lado a situaciones de fraudes o robos y por otro lado, a errores de facturación o de medición.

Este proyecto propone un modelo para predecir potenciales situaciones de casos de hurto basado en aprender el comportamiento de clientes que anteriormente hurtaron utilizando en la aprensión de los comportamientos las redes neuronales artificiales, de tal forma que la sensibilidad (verdaderos positivos) y especificidad (verdaderos negativos) sea el mayor posible.

Fase 2: Entendimiento de datos

Esta fase pretende coleccionar los datos iniciales, desarrollar procedimientos para entender los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

En esta etapa consolidamos los datos recibidos por la empresa eléctrica, hacemos análisis descriptivos de los mismos, evaluamos su consistencia y como resultado entregamos datos confiables y seleccionamos las variables que apoyen en la predicción para que en una fase posterior se utilicen en la preparación de datos-

La empresa eléctrica proporcionó tres archivos para iniciar la investigación:

Tabla 15: Información Inicial para la Investigación

Nombre de Archivo	Tipo de Archivo	Descripción de Contenido
Base Consumos	Microsoft Access Database (dmb)	Contiene datos de clientes, datos de la medición de los consumos mensuales correspondientes a 25 meses desde Octubre del 2007 a Octubre 2009 1'145,467.
Valorizaciones	Microsoft Excel 2003 (txt)	Contiene datos de la historia de los clientes hurtadores con su fecha de inicio de hurto de energía, fecha de fin de energía, cantidad de energía recuperada.
Tabla General	Microsoft Excel 2003 (txt)	Contiene la descripción de los datos de la tabla Base Consumos.

Fuente: Datos entregados por la empresa de distribución eléctrica para inicial la tesis.




 Base Consumos	14/04/2010 12:38 ...	Microsoft Access ...	1,480,232 KB
 TablaGeneral	24/12/2009 10:58 a...	Microsoft Excel 97...	138 KB
 Valorizaciones	24/12/2009 10:58 a...	WinZip File	973 KB

Figura 20: Información Inicial para la Investigación

Elaboración: La autora

En esta figura, mostramos los datos a analizar y separamos del grupo las variables que apoyarán al objetivo de predecir los fraudes de energía eléctrica en clientes residenciales.

Datos de Cliente	Consumo Mes	Dias Fact Mes	Clave Fact Mes	Clave Lectura Mes	VALORIZACIONES
NumeroCuenta	Consumo_24_D	Dias_Fact_24_D	Clave_Fact_24_D	Clave_Lect_24_D	NumeroCuenta
NumeroServicio	Consumo_23_D	Dias_Fact_23_D	Clave_Fact_23_D	Clave_Lect_23_D	NumeroServicio
CodigoDistrito	Consumo_22_D	Dias_Fact_22_D	Clave_Fact_22_D	Clave_Lect_22_D	NumeroInspeccion
Distrito	Consumo_21_D	Dias_Fact_21_D	Clave_Fact_21_D	Clave_Lect_21_D	AñoExpediente
Sector	Consumo_20_D	Dias_Fact_20_D	Clave_Fact_20_D	Clave_Lect_20_D	NumeroExpediente
Zona	Consumo_19_D	Dias_Fact_19_D	Clave_Fact_19_D	Clave_Lect_19_D	IdFechaCreacion
CodigoActividadComercial	Consumo_18_D	Dias_Fact_18_D	Clave_Fact_18_D	Clave_Lect_18_D	IdFechaInicioRecupero
ActividadComercial	Consumo_17_D	Dias_Fact_17_D	Clave_Fact_17_D	Clave_Lect_17_D	IdFechaFinRecupero
Tarifa	Consumo_16_D	Dias_Fact_16_D	Clave_Fact_16_D	Clave_Lect_16_D	TotalEnergia
TipoRedElectrica	Consumo_15_D	Dias_Fact_15_D	Clave_Fact_15_D	Clave_Lect_15_D	TotalRecupero
Potencia	Consumo_14_D	Dias_Fact_14_D	Clave_Fact_14_D	Clave_Lect_14_D	TotalIntereses
EstadoServicio	Consumo_13_D	Dias_Fact_13_D	Clave_Fact_13_D	Clave_Lect_13_D	TotalMoras
EstadoCliente	Consumo_12_D	Dias_Fact_12_D	Clave_Fact_12_D	Clave_Lect_12_D	
Marca	Consumo_11_D	Dias_Fact_11_D	Clave_Fact_11_D	Clave_Lect_11_D	
Modelo	Consumo_10_D	Dias_Fact_10_D	Clave_Fact_10_D	Clave_Lect_10_D	
Fase	Consumo_9_D	Dias_Fact_9_D	Clave_Fact_9_D	Clave_Lect_9_D	
Factor	Consumo_8_D	Dias_Fact_8_D	Clave_Fact_8_D	Clave_Lect_8_D	
SET	Consumo_7_D	Dias_Fact_7_D	Clave_Fact_7_D	Clave_Lect_7_D	
Alimentador	Consumo_6_D	Dias_Fact_6_D	Clave_Fact_6_D	Clave_Lect_6_D	
SED	Consumo_5_D	Dias_Fact_5_D	Clave_Fact_5_D	Clave_Lect_5_D	
Llave	Consumo_4_D	Dias_Fact_4_D	Clave_Fact_4_D	Clave_Lect_4_D	
	Consumo_3_D	Dias_Fact_3_D	Clave_Fact_3_D	Clave_Lect_3_D	
	Consumo_2_D	Dias_Fact_2_D	Clave_Fact_2_D	Clave_Lect_2_D	
	Consumo_1_D	Dias_Fact_1_D	Clave_Fact_1_D	Clave_Lect_1_D	
	Consumo_0_D	Dias_Fact_0_D	Clave_Fact_0_D	Clave_Lect_0_D	

Variables relacionadas a la medición del consumo

Figura 21: Campos Iniciales a Investigar

Elaboración: La autora

Estructura de datos de la tabla Base Consumo

Tabla 16: Estructura de Datos Información Original BaseConsumos

Columna	Tipo de Dato	Descripción
NumeroCuenta	Texto	Indica el número de cuenta del cliente
NumeroServicio	Texto	Indica el número de servicio de la cuenta. Es igual al número de cuenta.
Nombre	Texto	Nombre del cliente
Direccion	Texto	Dirección del cliente
CodigoDistrito	Texto	Código del distrito de residencia del cliente
Distrito	Texto	Nombre del distrito de residencia del cliente
Sector	Texto	Sector en el que se encuentra ubicada la casa del cliente
Zona	Texto	Zona a la que pertenece el sector
Correlat	Entero	Valor correlativo
SucursalComercial	Texto	-
CodigoActividadComercial	Texto	Código de la actividad comercial a la que se dedica el cliente.
ActividadComercial	Texto	Nombre de la actividad comercial a la que se dedica el cliente
Tarifa	Texto	-
TipoRedElectrica	Texto	Tipo de red eléctrica a la que pertenece el cliente.
Potencia	Decimal	-
TipoCliente	Texto	Tipo de cliente
EstadoServicio	Texto	-
EstadoCliente	Texto	-
Fecha_Ult_Deteccion	Date	Fecha en que ha sido detectado como hurtador por última vez en todo su historial.
Codigolrregularidad_Ult_Deteccion	Texto	Código de Irregularidad de la última detección de hurto
Irregularidad_Ult_Deteccion	Texto	Nombre de la Irregularidad de la última detección de hurto
NumeroNotificacion_Ult_Deteccion	Entero	Numero de Notificación de la última detección de hurto
Año_NumeroExpediente	Texto	Año del número del expediente.
Medidor	Texto	Medidor que tiene instalado el cliente.
Marca	Texto	Marca del medidor
Modelo	Texto	Modelo del medidor
Fase	Decimal	-
Factor	Decimal	-
[Estado Medidor]	Texto	Estado en que se encuentra el

		medidor.
FecInstalacion	Date	Fecha de instalación del medidor.
[Sello Medidor]	Texto	-
[Sello Bornera]	Texto	-
[Sello Contraste]	Texto	-
[SET]	Texto	-
Alimentador	Texto	-
SED	Texto	-
Llave	Texto	-
FecLectura_i	Date	Fecha de lectura del mes "i". Se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Correspondiente al consumo del cliente
[Dias Fact_i]	Entero	Días facturados del mes "i". Se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Correspondiente al consumo del cliente
[Clave Lect_i]	Texto	Clave de lectura del mes "i". Se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Correspondiente al consumo del cliente
[Clave Fact_i]	Texto	Clave de facturación del cliente para el mes número "i"
Consumo_i	Decimal	Consumo del mes "i" se repite 25 veces. Desde $i \geq 0$ e $i \leq 24$. Medido en Kw

Elaboración: La autora

Estructura de Datos de la tabla Valorizaciones

Tabla 17: Estructura de Datos de la Información a Investigar Valorizaciones

Columna	Tipo de Dato	Descripción
NumeroCuenta	Texto	Indica el número de cuenta del cliente
NumeroServicio	Texto	Indica el número de servicio de la cuenta. Es igual al número de cuenta.
NumeroInspeccion	Texto	Indica el número de la inspección realizada
AñoExpediente	Texto	Indica el año del expediente.
NumeroExpediente	Texto	Indica el número del expediente.
FechaCreacion	Texto	Señala la fecha de creación de la valorización
FechaInicioRecupero	Texto	Indica la fecha de inicio del recupero. La fecha en que comenzó a hurtar
FechaFinRecupero	Texto	Indica la fecha de fin de recupero. La fecha en que ya se controló el problema y ya no está hurtando.
TotalEnergia	Texto	Señala el total de energía hurtada.
TotalRecupero	Texto	Señala el total recuperado.
TotalIntereses	Texto	-
TotalMoras	Texto	-

Elaboración: La autora

Técnicas para el Entendimiento de Datos

Los procedimientos a realizar en esta etapa nos apoyamos en lo sugerido por Kasperu (2010) donde indica que las tareas a realizarse en esta etapa son:

- Examinar las propiedades gruesas de los datos.
- Estadística descriptiva Univariada y Multivariada.

Donde el resultado será:

- La visualización de datos.
- Fuente de los datos y frecuencia de adquisición.
- Formato de registro de los datos
- Influencia de la medición (frecuencia).
- Cantidad de datos (atributos y registros).
- Valores no permitidos.

Análisis univariado

Variables relacionadas con el consumo

Análisis de la clave de lectura

La clave de lectura es el estado observado en el que se encuentra el medidor al momento de realizar una lectura de energía.

En el cuadro, se presentan los valores posibles de la clave de lectura y su descripción.

000	Claves de Lectura
10	LECTURA NORMAL
20	MEDIDOR INTERNO - NO SE PUEDE LEER
21	CLIENTE NO PERMITE
22	CUBIERTOS POR MATERIALES/PLANTAS
23	FALTA LLAVE
24	MEDIDOR EN SUB. ESTACION
25	MEDIDOR INTERNO/COORDINAR LECTURA
40	NO EXISTE EL MEDIDOR
50	NO SE UBICA EL MEDIDOR
51	DIRECCION INEXACTA
60	MEDIDOR DETERIORADO
61	SIN CHAPA/MALOGRADA
62	SIN TAPA
63	SIN MICA
64	SERVICIO ELECTRIZADO
70	MEDIDOR CAMBIADO
80	MEDIDOR MANIPULADO
81	DA CORRIENTE A OTRO PREDIO (REVENTA)
90	MEDIDOR CON CONEXION INDEBIDA

Figura 22: Diccionario de Claves de Lectura

Elaboración: La autora

En la figura, se presenta el histograma de las claves de lectura de los 25 meses de análisis, donde se muestran los porcentajes más resaltantes.

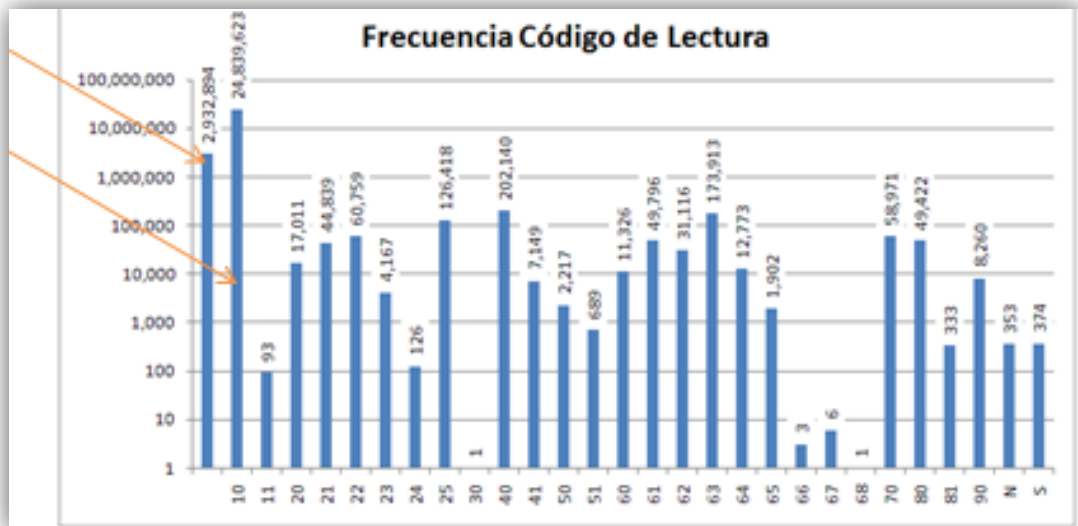


Figura 23: Histograma frecuencia Códigos de Lectura

Elaboración: La autora

Observación de la variable

- Existe un mayor porcentaje de casos de Código de lectura igual a 10 (86.7 %).
- Los códigos de lectura null (10.23 %) corresponde a los casos donde no se ha tenido consumo en ese mes.

Conclusión de la variable

Este atributo no se tomará en consideración para pre procesar los consumos mensuales, dado que su valor ha implicado una interpretación de un técnico respecto a la situación observada.

Análisis de la Clave de Facturación

La clave de facturación es el código que se le asigna según el tipo de facturación efectuado al cierre, puede ser facturado de manera normal, manual o promediado.

Tipos de clave de facturación:

- Normal (N)→ Facturación en la fecha.
- Promediado (*)→ Facturación cuando existe consumo, pero por diversos motivos, no se puede medir el consumo real en ese mes, el consumo se promediara.
- Manual (T)→ Facturación manual estimado, luego se convierte a N.
- Cerrado (U)→ Facturación cuando el consumo es cero para casos en que cancela el servicio al cliente debido a que no se puede medir el consumo en el tiempo establecido.

En la figura se presenta el histograma de la frecuencia de los claves de facturación, donde podrá se puede observar:

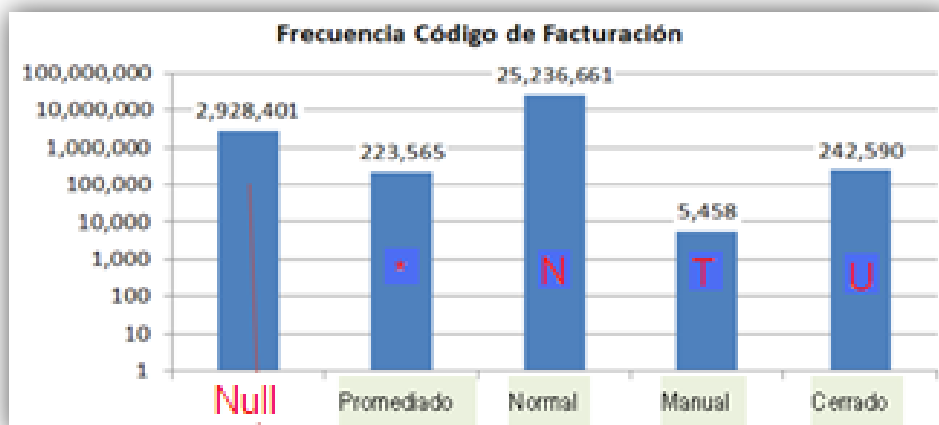


Figura 24: Histograma frecuencia Claves de Facturación

Elaboración: La autora

Observación de la variable

- La clave de facturación null corresponde a clientes nuevos.
- Existe un mayor porcentaje de claves de facturación de código normal.

Conclusión de la variable

Esta variable es importante porque ayudará a calcular el verdadero consumo del mes para los casos de consumo promediado o tipo asterisco que se deben a mediciones no realizadas y se encuentran acumuladas en el siguiente periodo.

Esta variable apoyará para la creación de un nuevo campo Código de Facturación.

Para facilitar el análisis de la clave de facturación de los 25 periodos de consumos, se procedió a crear la columna código de facturación conformado por todas claves de facturación (N***N**NNNN) de los 25 meses de consumo.

- Los vacíos a la izquierda indican clientes nuevos, es decir, tiene consumos solo en los últimos meses.
- No se ha encontrado casos donde exista un valor null entre dos claves de facturación.

- Si el código de facturación (CF) está formado por asteriscos a la izquierda seguido por una clave de tipo normal (N), y el CF (**N) entonces aplicar consumo promediado por mes.

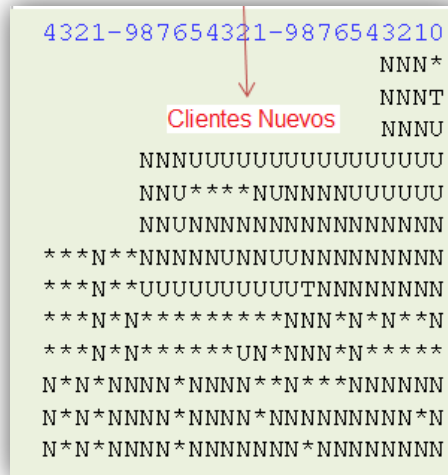


Figura 25: Análisis de clave de facturación

Elaboración: La autora

Análisis de los consumos

Esta columna es importante, dado que registra el valor del consumo medido en kilowatts (Kw).

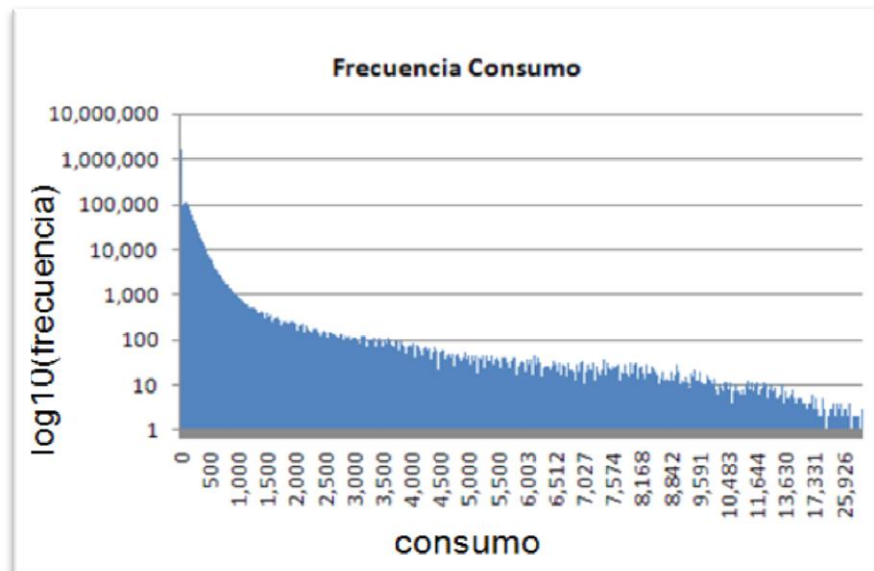


Figura 26: Análisis del consumo en kilowatts

Elaboración: La autora

Observación de la variable original

Se puede concluir

- El consumo tiene valores entre 0 y los 25,926 además de valores nulls.
- La distribución de frecuencia con valores originales informa muy poco, dado que los rangos son muy amplios.
- El consumo = null, se presenta por que el cliente recién se ha incorporado a la empresa y no presenta consumos.

Para facilitar el análisis, los datos observados se transformaran estos consumos aplicando la función $\log_{10}()$.

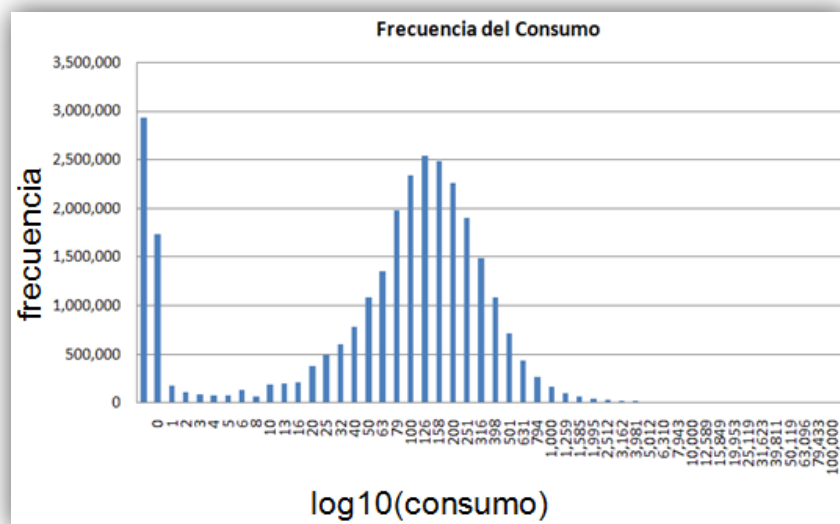


Figura 27: Histograma muestra consumo promedio logarítmico

Elaboración: La autora

En este histograma, se puede observar que el consumo promedio logarítmico es 126 Kw.

Análisis días de facturación

Esta columna es importante porque registra los días de facturación correspondiente a cada cliente por cada mes. Se extrajo el histograma de las frecuencias de los 25 meses de los días de facturación de todos los clientes.

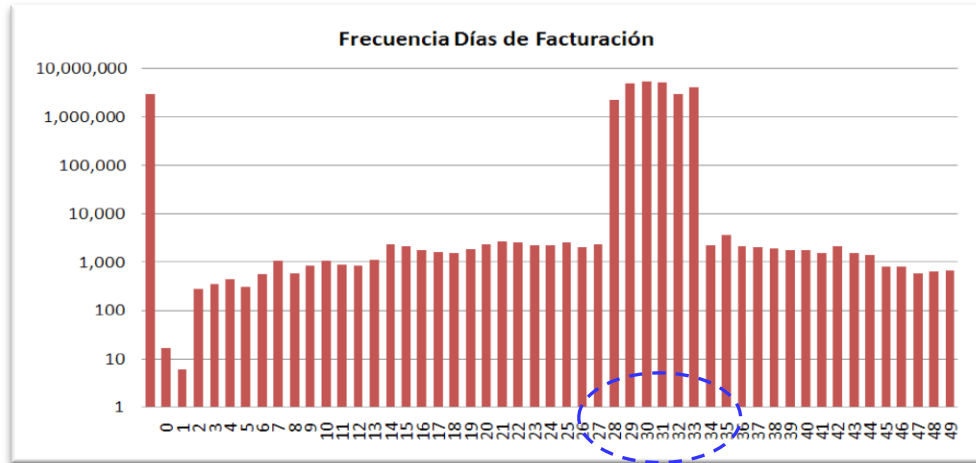


Figura 28: Frecuencia de días facturados

Elaboración: La autora

Observación de la variable original

- Los días de facturación se acumulan fuertemente entre 28 y 33 días.
- Existen 10.23% de valores null, debido a que tienen registrado consumos en menos de 25 meses.
- Los días de facturación mayores de 33 son días acumulados por casos de que no se puede hacer la medición.

Conclusión de la variable

Se debe reemplazar los días de facturación con los días obtenidos restando las fechas de lectura.

Fase 3: Preparación de Datos

Esta fase cubre todas las actividades necesarias para construir el conjunto de datos que serán provistos a las herramientas de modelado desde los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescrito. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Según los resultados obtenidos de la fase anterior, procederemos a hacer el proceso de extracción, limpieza, transformación y preparación de datos

Importación de datos

Las tablas originales son cargadas directamente a unas tablas temporales iniciales denominadas BaseConsumoPre y ValorizacionPre son las tablas encargadas de recibir los datos iniciales y almacenarlos en las tablas BaseConsumo y Valorizacion, a través de los procedimientos almacenados sp_Carga_Bases_Consumos y sp_Carga_Valorizacion. Se extraerán los datos desde las tablas BaseConsumoPre y ValorizacionPre para convertirlos a las estructuras establecidas de las tablas BaseConsumo y Valorizacion.

Creación de variables

- Código de facturación: Concatenación de claves de facturación de los 25 consumos. Esta variable es importante debido que luego con su estructura se calcularán los casos de consumo promediado, se eliminará registros para los casos tipo de facturación cerrada (U), eliminarán registros de cliente con menos de 12 meses de alta y también me servirá para el cálculo del consumo promedio diario. Ejemplo : N*****NT*NTUU

123456789-123456789-12345	len	Base	*	*N	rango	len	resto	pINI	pFIN	Base
*****N**NNNN**N**N	25	1	1	5	*****N	6	N**NNNNN**N**N**N	1	6	6
N**NNNN**N**N**N	20	6	2	2	*N	2	N**NNNN**N**N**N	7	8	8
N**NNNN**N**N**N	18	8	2	4	**N	4	NNNN**N**N**N	9	12	12
NNNN**N**N**N	14	12	5	5	*N	2	N**N**N	16	17	17
N**N**N	9	17	2	2	*N	2	NN**N	18	19	19
NN**N	7	19	3	4	**N	3	N*N	21	23	23
N*N	3	23	2	2	*N	2		24	25	25

Figura 29: Estructura de variable código de facturación.

Elaboración: La autora

- Promedio consumo: Esta variable es el resultado de la sumatoria de por cada cliente de los 25 meses dividido entre el tamaño real del código de facturación. Esta variable es importante porque más adelante servirá para eliminar los casos que tengan el consumo menores a 1 KW.

Transformación y cálculos de datos

- Claves de facturación: Los tipos de clave de facturación de tipo T se reemplazarán por el tipo N debido a que el especialista indicó que corresponden a mediciones manuales y corresponde a una medición correcta.
- Días de facturación: Los días de facturación según el análisis de datos oscila entre los días 28 y 33 días, se analizó que contiene días acumulados mas no los correspondientes al mes facturado, se evidenció que al restar las fechas de lectura del mes contra el mes anterior se obtiene los días de facturación real.

Mes	Clave Facturacion	Fecha Lectura	Días Facturados Originales	Días Facturados
24	*	06/10/2007	29	29
23	*	08/11/2007	62	33
22	*	07/12/2007	91	29
21	*	08/01/2008	123	32
20	*	07/02/2008	153	30
19	N	07/03/2008	182	29
18	*	07/04/2008	31	31
17	N	08/05/2008	62	31
16	*	07/06/2008	30	30
15	*	07/07/2008	60	30
14	*	07/08/2008	91	31
13	N	08/09/2008	123	32
12	N	09/10/2008	31	31
11	N	08/11/2008	30	30
10	N	06/12/2008	28	28
9	*	08/01/2009	33	33
8	N	07/02/2009	63	30
7	*	09/03/2009	30	30
6	N	07/04/2009	59	29
5	N	09/05/2009	32	32
4	*	08/06/2009	30	30
3	*	07/07/2009	59	29
2	N	08/08/2009	91	32
1	*	08/09/2009	31	31
0	N	09/10/2009	62	31

Figura 30: Días de facturación real y calculada.

Elaboración: La autora

De la figura, se puede, se visualiza que el día de facturación real para el periodo 15 es la diferencia de la fecha de lectura del periodo 15 menos la fecha de lectura del periodo 16.

Tablas a Utilizar en el proceso de ETL

A continuación, se describen las tablas temporales a utilizarse en la preparación de datos.

Tabla 18: Tablas a utilizar en el proceso ETL para la Preparación de Datos

Nombre de la Tablas Intermedias	Etapa	Descripción de la Tabla
1. BaseConsumoPre	Carga de datos	En esta tabla se importa los consumos tal como están en la base de datos original para permitir la carga directa. Todas sus columnas son de tipo nvarchar(255).
2. ValorizaciónPre	Carga de datos	En esta tabla se importa los registros con casos de hurto tal como están en la base de datos original, como todas sus columnas son de tipo varchar para permitir la carga directa.
3. BaseConsumo	Pre Procesamiento	Esta tabla contiene los consumos ya convertidos al tipo de dato necesario para su análisis. A la tabla BaseConsumo se aplico algunos procesos para permitir su carga, como por ejemplo las Fechas de última detección mayores que la fechas del análisis establecido en la tabla periodo actual.
4. Valorización	Pre Procesamiento	Esta tabla se carga desde la tabla valorizacionPre, donde las columnas son convertidas a sus tipos de datos correspondientes y filtrados por la fecha fin recuperación que pertenezcan al rango permitido establecido en la tabla Periodo actual.
5. Consumo	Pre Procesamiento	Esta tabla tiene la misma estructura que la tabla BaseConsumo pero con consumos promediados. Debido a que la tabla BaseConsumo contiene una corrección a los consumos mensuales en los casos en que haya consumos acumulados y luego distribuyendo el consumo del cliente.
6. Maestro	Pre Procesamiento	Esta tabla contienen los datos que han pasado por un filtro de acuerdo a ciertos parámetros establecidos. Los consumos en esta tabla son consumos promedio diario. (Clientes con estado Habilitado y tarifa BT5 B)

7. PeriodoActual	Pre procesamiento	En esta tabla se almacenaron las fechas de análisis para cada tipo de proceso Entrenamiento y Despliegue.
8. FiltroMaestro	Pre Procesamiento	En esta tabla se crea almacenaron los criterios para filtrar la tabla Consumo y producir la tabla Maestro.
9. Consulta	Modelado	Esta tabla se crea utilizando la tabla Maestro y la tabla Valorización se selecciona todos los casos de hurtos y los casos de no hurtos, servirá para la creación del modelo base.

Elaboración: La autora

En la siguiente sección, se presenta el detalle de cada una de las columnas de las tablas indicadas en el cuadro anterior.

Diccionario de Datos

Tabla BaseConsumoPre

Tabla 19: Diccionario de Datos de la Tabla BaseConsumoPre

Columna	Tipo de Dato	Default	Descripción
NumeroCuenta	nvarchar(255)	null	Indica el número de cuenta del cliente
NumeroServicio	nvarchar(255)	null	Indica el número de servicio de la cuenta. Es igual al número de cuenta.
Nombre	nvarchar(255)	null	Nombre del cliente
Direccion	nvarchar(255)	null	Dirección del cliente
CodigoDistrito	nvarchar(255)	null	código del distrito de residencia del cliente
Distrito	nvarchar(255)	null	Nombre del Distrito de residencia del cliente
Sector	nvarchar(255)	null	Sector en el que se encuentra ubicada la casa del cliente
Zona	nvarchar(255)	null	Zona a la que pertenece el sector
Correlat	nvarchar(255)	null	-
SucursalComercial	nvarchar(255)	null	-
CodigoActividadComercial	nvarchar(255)	null	Código de la actividad comercial a la que se dedica el cliente.
ActividadComercial	nvarchar(255)	null	Nombre de la actividad comercial a la que se dedica el cliente
Tarifa	nvarchar(255)	null	-
TipoRedElectrica	nvarchar(255)	null	Tipo de red eléctrica a la que pertenece el cliente.
Potencia	nvarchar(255)	null	-
TipoCliente	nvarchar(255)	null	Tipo de cliente

)		
EstadoServicio	nvarchar(255)	null	-
EstadoCliente	nvarchar(255)	null	-
Fecha_Ult_Deteccion	nvarchar(255)	null	Fecha en que ha sido detectado como hurtador por última vez en todo su historial.
CodigoIrregularidad_Ult_Deteccion	nvarchar(255)	null	código de irregularidad de la última detección de hurto
Irregularidad_Ult_Deteccion	nvarchar(255)	null	Nombre de la Irregularidad de la última detección de hurto
NumeroNotificacion_Ult_Deteccion	nvarchar(255)	null	Numero de Notificación de la última detección de hurto
Año_NumeroExpediente	nvarchar(255)	null	Año del número del expediente.
Medidor	nvarchar(255)	null	Medidor que tiene instalado el cliente.
Marca	nvarchar(255)	null	Marca del medidor
Modelo	nvarchar(255)	null	Modelo del medidor
Fase	nvarchar(255)	null	-
Factor	nvarchar(255)	null	-
[Estado Medidor]	nvarchar(255)	null	Estado en que se encuentra el medidor.
FecInstalacion	nvarchar(255)	null	Fecha de instalación del medidor.
[Sello Medidor]	nvarchar(255)	null	-
[Sello Bornera]	nvarchar(255)	null	-
[Sello Contraste]	nvarchar(255)	null	-
[SET]	nvarchar(255)	null	-
Alimentador	nvarchar(255)	null	-
SED	nvarchar(255)	null	-
Llave	nvarchar(255)	null	-
FecLectura_i	nvarchar(255)	null	Fecha de lectura del mes "i". Se repite 25 veces. Desde i>=0 e i<=24. Correspondiente al consumo del cliente
[Dias Fact_i]	nvarchar(255)	null	Días facturados del mes "i". Se repite 25 veces. Desde i>=0 e i<=24. Correspondiente al consumo del cliente
[Clave Lect_i]	nvarchar(255)	null	Clave de lectura del mes "i". Se repite 25 veces. Desde i>=0 e i<=24. Correspondiente al consumo del cliente
[Clave Fact_i]	nvarchar(255)	null	Clave de facturación del cliente para el mes número "i"
Consumo_i	nvarchar(255)	null	Consumo del mes "i" se repite 25 veces. Desde i>=0 e i<=24. Medido en Kw

Elaboración: La autora

Tabla ValorizacionPre

Tabla 20: Diccionario de datos de la Tabla ValorizacionPre

Columna	Tipo de Dato	Default	Descripción
NumeroCuenta	nvarchar(255)	null	Indica el número de cuenta del cliente
NumeroServicio	nvarchar(255)	null	Indica el número de servicio de la cuenta. Es igual al número de cuenta.
NumeroInspeccion	nvarchar(255)	null	Indica el número de la inspección realizada
AñoExpediente	nvarchar(255)	null	Indica el año del expediente.
NumeroExpediente	nvarchar(255)	null	Indica el número del expediente.
FechaCreacion	nvarchar(255)	null	Señala la fecha de creación de la valorización
FechaInicioRecupero	nvarchar(255)	null	Indica la fecha de inicio del recupero. La fecha en que comenzó a hurtar
FechaFinRecupero	nvarchar(255)	null	Indica la fecha de fin de recupero. La fecha en que ya se controló el problema y ya no está hurtando.
TotalEnergia	nvarchar(255)	null	Señala el total de energía hurtada.
TotalRecupero	nvarchar(255)	null	Señala el total recuperado.
TotalIntereses	nvarchar(255)	null	-
TotalMoras	nvarchar(255)	null	-

Elaboración: La autora

Tabla BaseConsumo

Tabla 21: Diccionario de Datos de la Tabla BaseConsumos

Columna	Pk	fk	Tipo de Dato	Default	Descripción
NumeroCuenta	x		int		Indica el número de cuenta del cliente
NumeroServicio	x		int		Indica el número de servicio de la cuenta. Es igual al número de cuenta.
Nombre			nvarchar(255)	null	Nombre del cliente
Direccion			nvarchar(255)	null	Dirección del cliente
CodigoDistrito			int	null	código del distrito de residencia del cliente
Distrito			nvarchar(255)	null	Nombre del distrito de residencia del cliente
Sector			int	null	Sector en el que se encuentra ubicada la casa del cliente
Zona			int	null	Zona a la que pertenece el sector

Correlat		int	null	-
SucursalComercial		int	null	-
CodigoActividadComercial		nvarchar(255)	null	Código de la actividad comercial a la que se dedica el cliente.
ActividadComercial		nvarchar(255)	null	Nombre de la actividad comercial a la que se dedica el cliente
Tarifa		nvarchar(255)	null	-
TipoRedElectrica		nvarchar(255)	null	Tipo de red eléctrica a la que pertenece el cliente.
Potencia		float	null	-
TipoCliente		nvarchar(255)	null	Tipo de cliente
EstadoServicio		int	null	-
EstadoCliente		nvarchar(255)	null	-
Fecha_Ult_Deteccion		datetime	null	Fecha en que ha sido detectado como hurtador por última vez en todo su historial.
CodigoIrregularidad_Ult_Deteccion		nvarchar(255)	null	código de irregularidad de la última detección de hurto
Irregularidad_Ult_Deteccion		nvarchar(255)	null	Nombre de la Irregularidad de la última detección de hurto
NumeroNotificacion_Ult_Deteccion		int	null	Numero de Notificación de la última detección de hurto
Año_NumeroExpediente		nvarchar(MAX)	null	Año del número del expediente.
Medidor		int	null	Medidor que tiene instalado el cliente.
Marca		nvarchar(255)	null	Marca del medidor
Modelo		nvarchar(255)	null	Modelo del medidor
Fase		nvarchar(255)	null	-
Factor		int	null	-
[Estado Medidor]		nvarchar(255)	null	Estado en que se encuentra el medidor.
Feclnstalacion		datetime	null	Fecha de instalación del medidor.
[Sello Medidor]		nvarchar(255)	null	-
[Sello Bornera]		nvarchar(255)	null	-
[Sello Contraste]		nvarchar(255)	null	-

[SET]			nvarchar(255)	null	-
Alimentador			nvarchar(50)	null	-
SED			nvarchar(255)	null	-
Llave			int	null	-
FecLectura_i			datetime	null	Fecha de lectura del mes número "i" del cliente. Para nuestro caso se nos dio la tabla desde i>=0 e i<=24
[Días Fact_i]			int	null	Días facturados en el mes número "i" del cliente
[Clave Lect_i]			nvarchar(255)	null	Clave de lectura del cliente para el mes número "i"
[Clave Fact_i]			nvarchar(255)	null	Clave de facturación del cliente para el mes número "i"
Consumo_i			int	null	Consumo en el mes número "i" en kw

Elaboración: La autora

Tabla Valorización

Tabla 22: Diccionario de Datos de la Tabla Valorizacion

Columna	P k	F k	Tipo de Dato	Defaul t	Descripción
NumeroCuenta		x	int		Indica el número de cuenta del cliente
NumeroServicio		x	int		Indica el número de servicio de la cuenta. Es igual al número de cuenta.
NumeroInspeccion	x		int		Indica el número de la inspección realizada
AñoExpediente			int	null	Indica el año del expediente.
NumeroExpediente			int	null	Indica el número del expediente.
FechaCreacion			datetime	null	Señala la fecha de creación de la valorización
FechaInicioRecupero			datetime	null	Indica la fecha de inicio del recupero. La fecha en que comenzó a hurtar
FechaFinRecupero			datetime	null	Indica la fecha de fin de recupero. La fecha en que ya se controló el problema y ya no está hurtando.
TotalEnergia			float	null	Señala el total de energía hurtada.
TotalRecupero			float	null	Señala el total recuperado.
TotalIntereses			float	null	-
TotalMoras			float	null	-

Elaboración: La autora

Tabla Consumo

Tabla 23: Diccionario de Datos de la Tabla Consumos

Columna	Pk	Fk	Tipo de Dato	Default	Descripción
NumeroCuenta	x		int		Indica el número de cuenta del cliente
NumeroServicio	x		int		Indica el número de servicio de la cuenta. Es igual al número de cuenta.
Nombre			nvarchar(255)	null	Nombre del cliente
...
...
[Clave Fact_0]			nvarchar(255)	null	Clave de facturación del cliente para el mes 09/2009
Consumo_0			int	null	Consumo en el mes 09/2009 en kw
CodigoFacturacion			nvarchar(255)	null	Indica los periodos de facturación concatenados: NNNNNN*NNNNNNNNNNNNNN NNNNN
PromedioConsumo			int	null	Indica el consumo promedio

Elaboración: La autora

Tabla Maestro

Tabla 24: Diccionario de Datos de la Tabla Maestro

Columna	Pk	fk	Tipo de Dato	Default	Descripción
NumeroCuenta	x		int		Indica el número de cuenta del cliente
NumeroServicio	x		int		Indica el número de servicio de la cuenta. Es igual al número de cuenta.
CodigoDistrito			int	null	código del distrito de residencia del cliente
Distrito			nvarchar(255)	null	Nombre del distrito de residencia del cliente
Sector			int	null	Sector en el que se encuentra ubicada la casa del cliente
Zona			int	null	Zona a la que pertenece el sector
CodigoActividadComercial			nvarchar(255)	null	Código de la actividad comercial a la que se dedica el cliente.
ActividadComercial			nvarchar(255)	null	Nombre de la actividad comercial a la que se dedica el cliente
Tarifa			nvarchar(255)	null	-
TipoRedElectrica			nvarchar(255)	null	Tipo de red eléctrica a la que pertenece el cliente.
Potencia			float	null	-

EstadoServicio			int	null	-
EstadoCliente			nvarchar(255)	null	-
Marca			nvarchar(255)	null	Marca del medidor
Modelo			nvarchar(255)	null	Modelo del medidor
Fase			nvarchar(255)	null	-
Factor			int	null	-
[SET]			nvarchar(255)	null	-
Alimentador			nvarchar(50)	null	-
SED			nvarchar(255)	null	-
Llave			int	null	-
HurtoMeses			int	null	Cantidad de meses en los que ha hurtado
Hurtold			varchar(255)	null	Indica 'si' si es que ha hurtado alguna vez en los últimos 25 meses.
Consumo_i			decimal(20, 4)	0	Consumo en el mes "i" en kw. Tenemos 25 columnas de este tipo, una por cada mes
Hurto_i			varchar(255)	no	Indica 'si' si es que ha hurtado en el mes "i". Tenemos 25 columnas de este tipo.

Elaboración: La autora

Tabla Periodo Actual

Tabla 25: Diccionario de Datos de la Tabla PeriodoActual

Columna	Pk	fk	Tipo de Dato	Default	Descripción
Fechalni			datetime	null	Indica la Fecha inicial a analizar considerar que siempre empiece por la primera semana de cada mes.
FechaFin			datetime	null	Indica la Fecha final de análisis considerar que siempre pertenezca a la última semana del mes.
PeriodoBase			varchar(50)	null	Cantidad de meses en los que ha hurtado.
Operacion			varchar(100)	null	Indica 'si' si es que ha hurtado alguna vez en los últimos 25 meses.
PrimeraDeteccion			datetime	null	Indica 'si' si el registro tiene un hurto en el último mes.

Elaboración: La autora

Tabla FiltroMaestro

Tabla 26: Diccionario de Datos de la Tabla FiltroMaestro

Columna	Pk	fk	Tipo de Dato	Default	Descripción
Columna			varchar(1000)	null	Indica la columna que se va a filtrar, para este caso 'Tarifa'
Valores			varchar(1000)	null	Indica el valor de la columna (BT5 B)
Estado			varchar(1000)	null	Indica el estado de la tarifa BT5 B , para este caso es el valor 1.

Elaboración: La autora

Tabla Consulta

Tabla 27: Diccionario de Datos de la Tabla Consulta

Columna	Pk	fk	Tipo de Dato	Default	Descripción
idT	x		int		Indica la numeración de los casos de hurto y no hurto que existen.
NumeroCuenta	x		int		Indica el número de cuenta del cliente
HurtoMeses			int	null	Cantidad de meses en los que ha hurtado
Hurtold			varchar(255)	null	Indica 'si' si es que ha hurtado alguna vez en los últimos 25 meses.
HurtoldK			varchar(255)	null	Indica 'si' si el registro tiene un hurto en el último mes.
Consumo_i			decimal(20, 4)	0	Consumo en el mes "i" en kw. Tenemos 25 columnas de este tipo, una por cada mes
Hurto_i			varchar(255)	no	Indica 'si' si es que ha hurtado en el mes "i". Tenemos 25 columnas de este tipo.

Elaboración: La autora

Importación de datos

BaseConsumoPre y ValorizacionPre son las tablas encargadas de recibir los datos iniciales y almacenarlos en las tablas BaseConsumo y Valorizacion, a través de los procedimientos almacenados **sp_Carga_Bases_Consumos** y **sp_Carga_Valorizacion**. Se extraerán los datos desde las tablas BaseConsumoPre y ValorizacionPre para convertirlos a las estructuras establecidas de las tablas BaseConsumo y Valorizacion.

Paso para importar datos originales

1.- Crear las tablas BaseConsumoPre y ValorizacionPre todas con tipo de dato nvarchar (255) para no perder datos en la carga luego en un segundo proceso se formateará los datos según corresponda.

Script de creación de tablas donde se cargarán fuentes originales.

```
USE BD_Taller_Tesis
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE [dbo].[BaseConsumoPre] (
    [NumeroCuenta] [nvarchar] (255) NULL,
    [NumeroServicio] [nvarchar] (255) NULL,
    [Nombre] [nvarchar] (255) NULL,
    [Direccion] [nvarchar] (255) NULL,
    [CodigoDistrito] [nvarchar] (255) NULL,
    [Distrito] [nvarchar] (255) NULL,
    [Sector] [nvarchar] (255) NULL,
    [Zona] [nvarchar] (255) NULL,
    [Correlat] [nvarchar] (255) NULL,
    [SucursalComercial] [nvarchar] (255) NULL,
    [CodigoActividadComercial] [nvarchar] (255) NULL,
    [ActividadComercial] [nvarchar] (255) NULL,
    [Tarifa] [nvarchar] (255) NULL,
    [TipoRedElectrica] [nvarchar] (255) NULL,
    [Potencia] [nvarchar] (255) NULL,
    [TipoCliente] [nvarchar] (255) NULL,
    [EstadoServicio] [nvarchar] (255) NULL,
    [EstadoCliente] [nvarchar] (255) NULL,
    [Fecha_Ult_Deteccion] [nvarchar] (255) NULL,
    [CodigoIrregularidad_Ult_Deteccion] [nvarchar] (255)
    NULL,
    [Irregularidad_Ult_Deteccion] [nvarchar] (255) NULL,
    [NumeroNotificacion_Ult_Deteccion] [nvarchar] (255) NULL,
    [Año_NumeroExpediente] [nvarchar] (255) NULL,
    [Medidor] [nvarchar] (255) NULL,
    [Marca] [nvarchar] (255) NULL,
    [Modelo] [nvarchar] (255) NULL,
    [Fase] [nvarchar] (255) NULL,
    [Factor] [nvarchar] (255) NULL,
    [Estado Medidor] [nvarchar] (255) NULL,
    [FecInstalacion] [nvarchar] (255) NULL,
    [Sello Medidor] [nvarchar] (255) NULL,
    [Sello Bornera] [nvarchar] (255) NULL,
    [Sello Contraste] [nvarchar] (255) NULL,
    [SET] [nvarchar] (255) NULL,
    [Alimentador] [nvarchar] (50) NULL,
    [SED] [nvarchar] (255) NULL,
    [Llave] [nvarchar] (255) NULL,
    [FecLectura_24] [nvarchar] (255) NULL,
    [Dias Fact_24] [nvarchar] (255) NULL,
    [Clave Lect_24] [nvarchar] (255) NULL,
    [Clave Fact_24] [nvarchar] (255) NULL,
    [Consumo_24] [nvarchar] (255) NULL,
```

[FecLectura_23] [nvarchar] (255) NULL,
[Dias Fact_23] [nvarchar] (255) NULL,
[Clave Lect_23] [nvarchar] (255) NULL,
[Clave Fact_23] [nvarchar] (255) NULL,
[Consumo_23] [nvarchar] (255) NULL,
[FecLectura_22] [nvarchar] (255) NULL,
[Dias Fact_22] [nvarchar] (255) NULL,
[Clave Lect_22] [nvarchar] (255) NULL,
[Clave Fact_22] [nvarchar] (255) NULL,
[Consumo_22] [nvarchar] (255) NULL,
[FecLectura_21] [datetime] NULL,
[Dias Fact_21] [nvarchar] (255) NULL,
[Clave Lect_21] [nvarchar] (255) NULL,
[Clave Fact_21] [nvarchar] (255) NULL,
[Consumo_21] [nvarchar] (255) NULL,
[FecLectura_20] [nvarchar] (255) NULL,
[Dias Fact_20] [nvarchar] (255) NULL,
[Clave Lect_20] [nvarchar] (255) NULL,
[Clave Fact_20] [nvarchar] (255) NULL,
[Consumo_20] [nvarchar] (255) NULL,
[FecLectura_19] [nvarchar] (255) NULL,
[Dias Fact_19] [nvarchar] (255) NULL,
[Clave Lect_19] [nvarchar] (255) NULL,
[Clave Fact_19] [nvarchar] (255) NULL,
[Consumo_19] [nvarchar] (255) NULL,
[FecLectura_18] [nvarchar] (255) NULL,
[Dias Fact_18] [nvarchar] (255) NULL,
[Clave Lect_18] [nvarchar] (255) NULL,
[Clave Fact_18] [nvarchar] (255) NULL,
[Consumo_18] [nvarchar] (255) NULL,
[FecLectura_17] [nvarchar] (255) NULL,
[Dias Fact_17] [nvarchar] (255) NULL,
[Clave Lect_17] [nvarchar] (255) NULL,
[Clave Fact_17] [nvarchar] (255) NULL,
[Consumo_17] [nvarchar] (255) NULL,
[FecLectura_16] [nvarchar] (255) NULL,
[Dias Fact_16] [nvarchar] (255) NULL,
[Clave Lect_16] [nvarchar] (255) NULL,
[Clave Fact_16] [nvarchar] (255) NULL,
[Consumo_16] [nvarchar] (255) NULL,
[FecLectura_15] [nvarchar] (255) NULL,
[Dias Fact_15] [nvarchar] (255) NULL,
[Clave Lect_15] [nvarchar] (255) NULL,
[Clave Fact_15] [nvarchar] (255) NULL,
[Consumo_15] [nvarchar] (255) NULL,
[FecLectura_14] [nvarchar] (255) NULL,
[Dias Fact_14] [nvarchar] (255) NULL,
[Clave Lect_14] [nvarchar] (255) NULL,
[Clave Fact_14] [nvarchar] (255) NULL,
[Consumo_14] [nvarchar] (255) NULL,
[FecLectura_13] [nvarchar] (255) NULL,
[Dias Fact_13] [nvarchar] (255) NULL,
[Clave Lect_13] [nvarchar] (255) NULL,
[Clave Fact_13] [nvarchar] (255) NULL,
[Consumo_13] [nvarchar] (255) NULL,
[FecLectura_12] [nvarchar] (255) NULL,
[Dias Fact_12] [nvarchar] (255) NULL,
[Clave Lect_12] [nvarchar] (255) NULL,
[Clave Fact_12] [nvarchar] (255) NULL,
[Consumo_12] [nvarchar] (255) NULL,
[FecLectura_11] [nvarchar] (255) NULL,

```

[Dias Fact_11] [nvarchar] (255) NULL,
[Clave Lect_11] [nvarchar] (255) NULL,
[Clave Fact_11] [nvarchar] (255) NULL,
[Consumo_11] [nvarchar] (255) NULL,
[FecLectura_10] [nvarchar] (255) NULL,
[Dias Fact_10] [int] NULL,
[Clave Lect_10] [nvarchar] (255) NULL,
[Clave Fact_10] [nvarchar] (255) NULL,
[Consumo_10] [nvarchar] (255) NULL,
[FecLectura_9] [nvarchar] (255) NULL,
[Dias Fact_9] [nvarchar] (255) NULL,
[Clave Lect_9] [nvarchar] (255) NULL,
[Clave Fact_9] [nvarchar] (255) NULL,
[Consumo_9] [nvarchar] (255) NULL,
[FecLectura_8] [nvarchar] (255) NULL,
[Dias Fact_8] [nvarchar] (255) NULL,
[Clave Lect_8] [nvarchar] (255) NULL,
[Clave Fact_8] [nvarchar] (255) NULL,
[Consumo_8] [nvarchar] (255) NULL,
[FecLectura_7] [nvarchar] (255) NULL,
[Dias Fact_7] [nvarchar] (255) NULL,
[Clave Lect_7] [nvarchar] (255) NULL,
[Clave Fact_7] [nvarchar] (255) NULL,
[Consumo_7] [nvarchar] (255) NULL,
[FecLectura_6] [nvarchar] (255) NULL,
[Dias Fact_6] [nvarchar] (255) NULL,
[Clave Lect_6] [nvarchar] (255) NULL,
[Clave Fact_6] [nvarchar] (255) NULL,
[Consumo_6] [nvarchar] (255) NULL,
[FecLectura_5] [nvarchar] (255) NULL,
[Dias Fact_5] [nvarchar] (255) NULL,
[Clave Lect_5] [nvarchar] (255) NULL,
[Clave Fact_5] [nvarchar] (255) NULL,
[Consumo_5] [nvarchar] (255) NULL,
[FecLectura_4] [nvarchar] (255) NULL,
[Dias Fact_4] [nvarchar] (255) NULL,
[Clave Lect_4] [nvarchar] (255) NULL,
[Clave Fact_4] [nvarchar] (255) NULL,
[Consumo_4] [nvarchar] (255) NULL,
[FecLectura_3] [nvarchar] (255) NULL,
[Dias Fact_3] [nvarchar] (255) NULL,
[Clave Lect_3] [nvarchar] (255) NULL,
[Clave Fact_3] [nvarchar] (255) NULL,
[Consumo_3] [nvarchar] (255) NULL,
[FecLectura_2] [nvarchar] (255) NULL,
[Dias Fact_2] [nvarchar] (255) NULL,
[Clave Lect_2] [nvarchar] (255) NULL,
[Clave Fact_2] [nvarchar] (255) NULL,
[Consumo_2] [nvarchar] (255) NULL,
[FecLectura_1] [nvarchar] (255) NULL,
[Dias Fact_1] [nvarchar] (255) NULL,
[Clave Lect_1] [nvarchar] (255) NULL,
[Clave Fact_1] [nvarchar] (255) NULL,
[Consumo_1] [nvarchar] (255) NULL,
[FecLectura_0] [nvarchar] (255) NULL,
[Dias Fact_0] [nvarchar] (255) NULL,
[Clave Lect_0] [nvarchar] (255) NULL,
[Clave Fact_0] [nvarchar] (255) NULL,
[Consumo_0] [nvarchar] (255) NULL
) ON [PRIMARY]

```



```

USE BD_Taller_Tesis
GO

SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE [dbo].[ValorizacionPre] (
    [NumeroCuenta] [nvarchar] (255) NULL,
    [NumeroServicio] [nvarchar] (255) NULL,
    [NumeroInspeccion] [nvarchar] (255) NULL,
    [AñoExpediente] [nvarchar] (255) NULL,
    [NumeroExpediente] [nvarchar] (255) NULL,
    [FechaCreacion] [nvarchar] (255) NULL,
    [FechaInicioRecupero] [nvarchar] (255) NULL,
    [FechaFinRecupero] [nvarchar] (255) NULL,
    [TotalEnergia] [nvarchar] (255) NULL,
    [TotalRecupero] [nvarchar] (255) NULL,
    [TotalIntereses] [nvarchar] (255) NULL,
    [TotalMoras] [nvarchar] (255) NULL
) ON [PRIMARY]

```

2.- Cargar las fuentes iniciales Ir a Task > Import Data

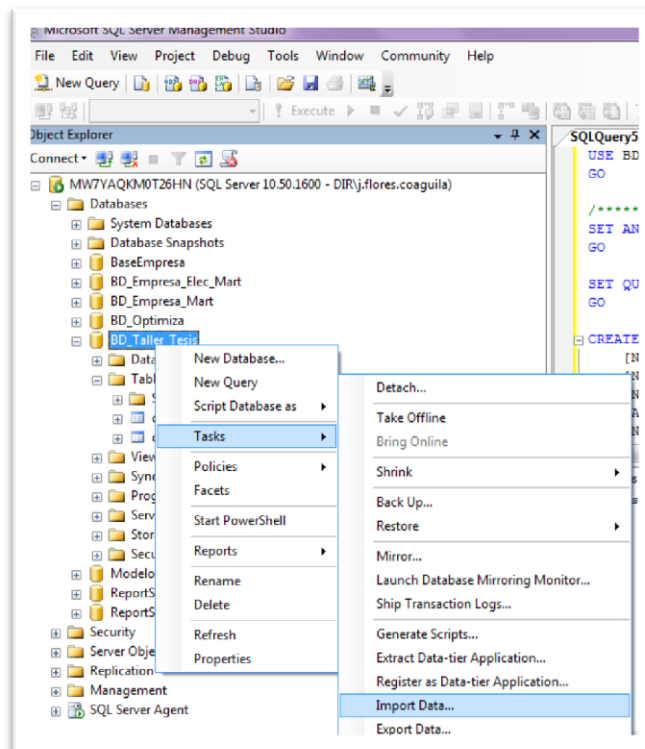


Figura 32: Importación de fuentes de datos originales.

Elaboración: La autora

- 3- Seleccionar el tipo de Origen de Datos, en este caso es un archivo de Microsoft Access Database.

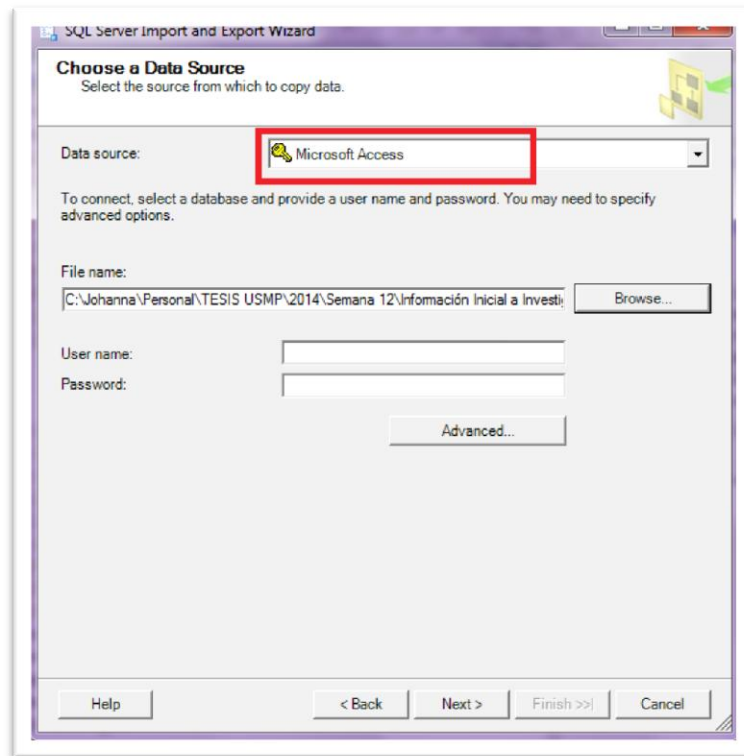


Figura 33: Selección de Tipo de Archivo de Origen de Datos.

Elaboración: La autora

- 4.- Cargar el archivo de Accsses .

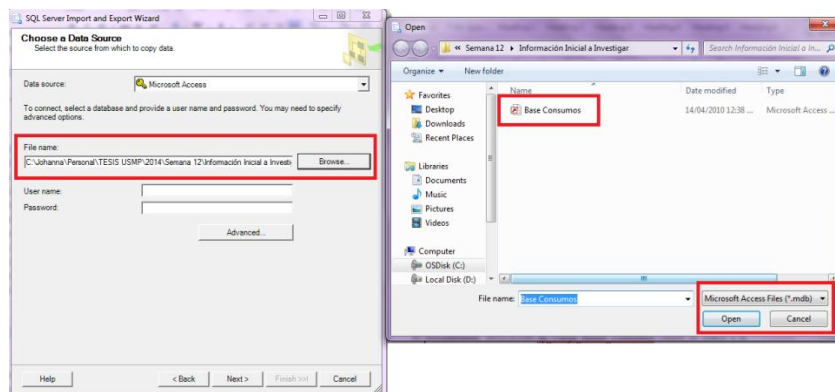


Figura 34: Selección de Origen de Destino de archivos originales.

Elaboración: La autora

5.- Seleccionar el destino y a que BD se cargará los datos.

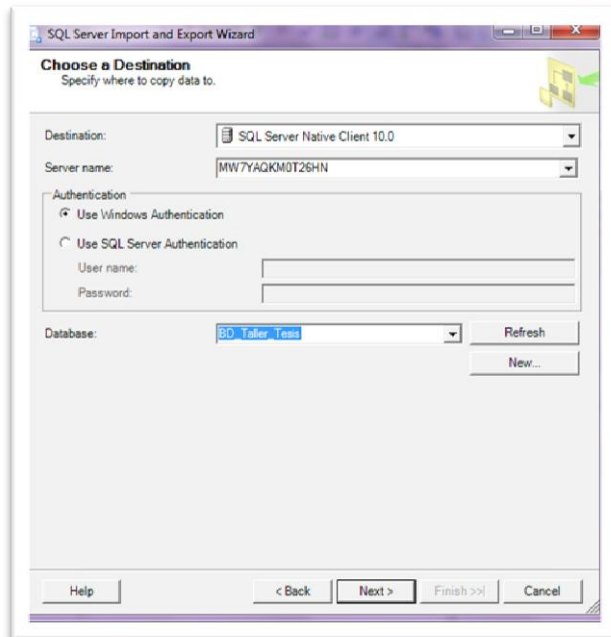


Figura 35: Selección de Base de Datos donde se cargarán las fuentes iniciales.

Elaboración: La autora

6.- Seleccionamos copiar datos uno o más tablas o vistas a la BD.

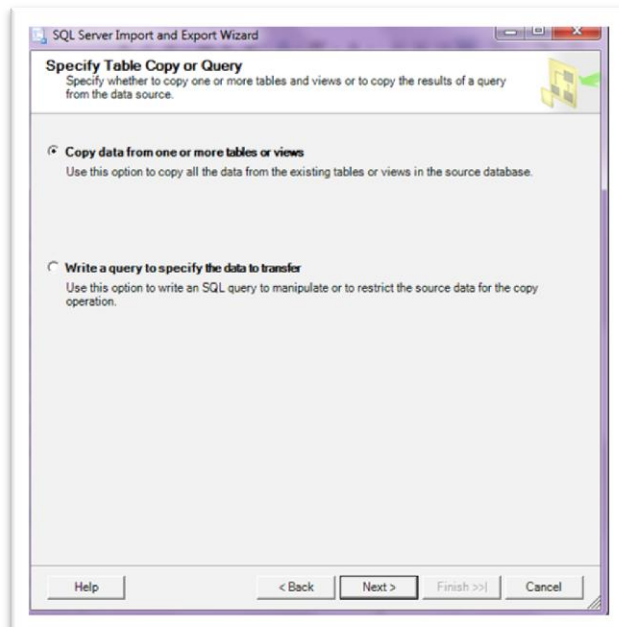


Figura 36: Selección de datos a copiar a la Base de Datos.

Elaboración: La autora

7.- Elegimos el origen y el nombre de la tabla destino creada en la BD.

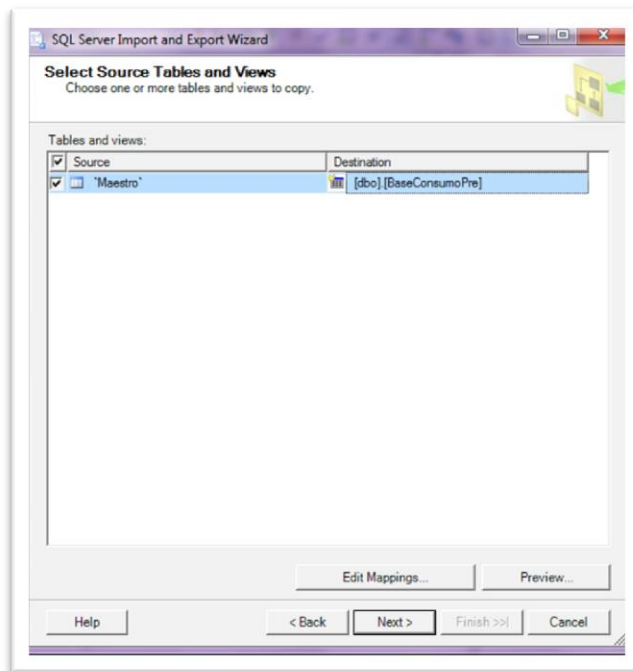


Figura 37: Selección de tabla destino en la Base de datos.

Elaboración: La autora

8.- Validamos que las columnas origen correspondan con las columnas destinos y se cargue completamente los datos.

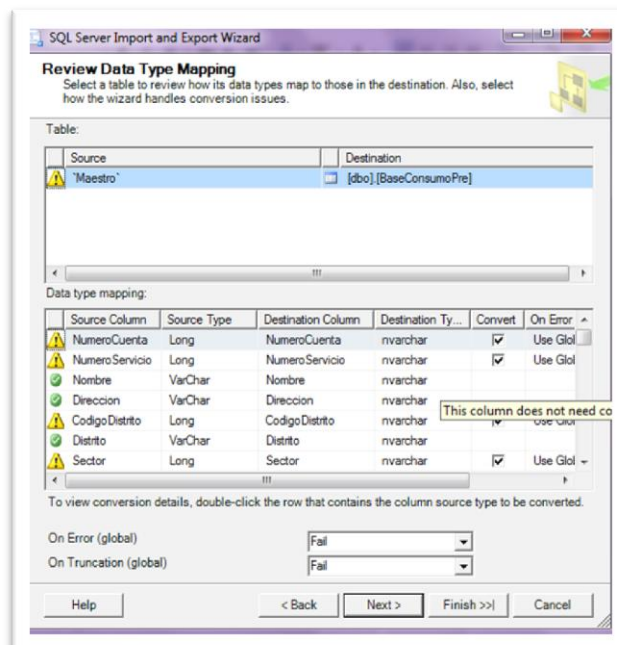


Figura 38: Seleccionar las columnas de origen y destino en la carga coincidan.

Elaboración: La autora

9.- Se carga todos los datos de la fuente original a la tabla destino.

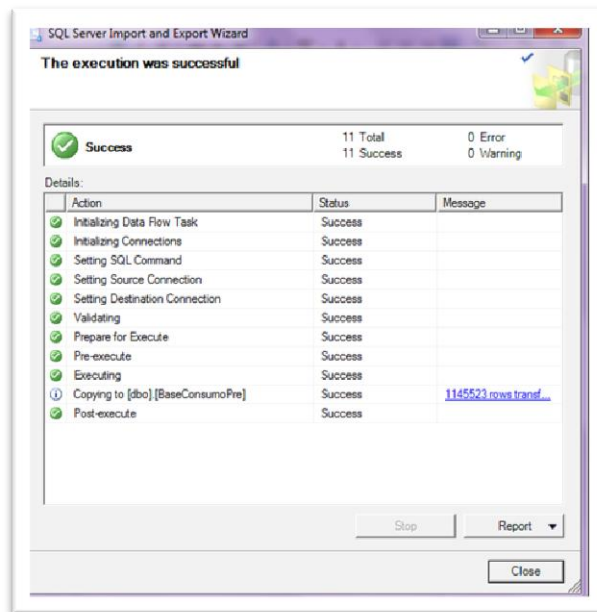


Figura 39: Validación de carga satisfactoria.

Elaboración: La autora

10.- Los mismos pasos seguimos para cargar los datos originales de Valorizaciones cambiamos el tipo de origen de datos en este caso sería de tipo txt.

Procedimientos desarrollados para la preparación de datos:

A continuación, se definirán todos los procedimientos que participan en la preparación inicial de Datos:



Figura 40: Secuencia de procedimiento de preparación de Datos

Elaboración: La autora

Procedimiento sp_0_base_total

Este es el procedimiento principal o base que inicia todo el proceso donde se configura los parámetros iniciales y se define los procedimientos almacenados que se utilizarán en la etapa de pre procesamiento.

Tener especial consideración para los valores de estos parámetros, debido a que es sensible por lo que su cambio debe ser informado al proveedor.

```
CREATE PROCEDURE [dbo].[sp_0_base_total] AS
BEGIN
declare      @FechaIni          datetime
declare      @FechaFin          datetime
declare      @PrimeraDeteccion  datetime
declare      @K                  int
declare      @PeriodoBase       varchar(20)
declare      @Operacion         varchar(100)
declare      @ConsumoMinimo     decimal(20,4)
declare      @DiasMax33         int

select      @K=12
select      @ConsumoMinimo=1.0
select      @DiasMax33=33

delete from Consumo
delete from Maestro
delete from Consulta

select      @FechaIni          = '10/01/2007'
select      @FechaFin          = '10/31/2009'
select      @PrimeraDeteccion = '01/01/1998'
select      @Operacion         = 'Entrenamiento'

delete from PeriodoActual
where Operacion = 'Entrenamiento'

exec sp_5fechaperiodo @FechaFin,@PeriodoBase output

insert Into PeriodoActual
values(@FechaIni ,@FechaFin,@PeriodoBase,
@Operacion,@PrimeraDeteccion)

exec sp_0_carga_total @ConsumoMinimo , @K ,@DiasMax33

END
```

Procedimiento sp_0_Carga_total

Este procedimiento es el segundo en el orden de ejecución es llamado por el procedimiento sp_base_total y utiliza los parámetros definidos inicialmente que servirán para realizar el pre procesamiento de datos, recibe los

parámetros de consumo mínimo, tamaño del vector de consumos y los días máximos de facturación.

```
CREATE PROCEDURE [dbo].[sp_0_carga_total]
(@ConsumoMinimo decimal(20,4), @K int ,@DiasMax33 int)as
begin
declare @FechaIni datetime
declare @FechaFin datetime

exec sp_0_Carga_Bases_Consumos
exec sp_0_Carga_ValORIZACION
exec sp_1_prepara_datos @ConsumoMinimo, @K, @DiasMax33

select @FechaIni=FechaIni
from PeriodoActual
where Operacion= 'Entrenamiento'

select @FechaFin =FechaFin
from PeriodoActual
where Operacion= 'Entrenamiento'

exec sp_2_carga_hurto_maestro @FechaIni,@FechaFin
exec sp_3_carga_hurto_casos @K

end
```

Definición de parámetros

Debe tener especial cuidado en su registro dado que todos los procedimientos involucrados hacen uso directo o indirecto de estos parámetros.

Por otro lado, debe tenerse especial cuidado en colocar la configuración regional del servidor y de las estaciones cliente en: Panel de Control → Configuración Regional y de Idiomas → Ubicación → seleccionar Perú verificar que los números para símbolos decimales se usa el punto (.), símbolos de separación de miles (,) y separador de listas (,).

Tabla 28: Descripción de Parámetros del Procedimiento
SP_0_CARGA_TOTAL

Parámetro	Tipo de Dato	Ejemplo de Valor	Descripción
@FechaIni	datetime	10/01/2007	Corresponde al intervalo de fechas desde las cuales se han obtenido los datos.
@FechaFin	Datetime	10/31/2009	
@PrimeraDeteccion	Datetime	01/01/1998	Fecha Mínima para el campo Fecha_Ult_Deteccion
@K	int	12	Periodo de análisis, en este caso 1 año.
@Operacion	varchar(20)	Entrenamiento	El parámetro de operación se usara según el proceso que se ejecute, puede ser de 'Entrenamiento' y 'Prueba'
@ConsumoMinimo	decimal(20,4)	1.0	Consumo mínimo permitido en Kw.
@DiasMax33	int	33	Este parámetro fue creado para analizar el día máximo permitido para el mes 25 (Dia fact_24)

Elaboración: La autora

Este procedimiento consta de cinco procesos internos:

1. sp_0_Carga_Bases_Consumos : Procedimiento que Carga información inicial.
2. sp_0_Carga_Valorizacion : Procedimiento que Carga información inicial.
3. sp_1_prepara_datos : Carga los datos preparados de tabla BaseConsumo a la tabla Consumo, prepara y limpia los registros que se encuentran en la tabla Consumo a través del procedimiento **sp_1_prepara_datos**.
4. sp_2_carga_hurto_maestro : Cargar de la tabla Consumo los clientes habilitados y con tarifa BT5 B a la tabla Maestro, consulta con la tabla valorización para asignar los hurtos mes a mes.
5. sp_3_carga_hurto_casos : Carga a través del procedimiento

sp_3_carga_hurto_Maestro.

Carga los casos de hurto de la tabla Maestro y los casos que nunca tuvieron hurto y colocarlos en la tabla Consulta y esta tabla que se utilizara para crear el modelo, a través del procedimiento **sp_3_carga_hurto_casos**.

Tabla 29: Descripción de Procedimiento SP_0_Carga_Total

NOMBRE STORE sp_0_carga_total		
PROCEDURE		
PARAMETROS QUE RECIBE	TIPO	
@ConsumoMinimo	decimal(20,4)	
@K	int	
@DiasMax33	int	
PARAMETROS DECLARADOS	TIPO	DESCRIPCION
@FechaIni	datetime	Guardara la fecha inicial de la tabla periodo actual para el tipo "Entrenamiento"
@FechaFin	Datetime	Guardara la fecha final de la tabla periodo actual para el tipo "Entrenamiento"
SUB STORE PROCEDURE	PARAMETROS RECIBE	DESCRIPCIÓN
➤ sp_0_Carga_Bases_Consumos		Este sub proceso se encarga de preparar la data para que pueda ser cargada a la tabla principal BaseConsumo
➤ sp_0_Carga_Valorizacion		Este sub proceso se encarga de preparar la data para que pueda ser cargada a la tabla principal Valorizacion
➤ sp_1_prepara_datos	@ConsumoMinimo	Este sub proceso se encarga de la preparación y limpieza de los datos.
	@K	
	@DiasMax33	
➤ sp_2_carga_hurto_maestro	@FechaIni	Este sub proceso carga los consumos promediado diarios y actualiza si existe un hurto en cada mes.
	@FechaFin	
➤ sp_3_carga_hurto_casos	@K	Este sub proceso se encarga del desfase para los casos de hurto.

Elaboración: La autora

Procedimiento sp_0_Carga_Bases_Consumos:

El procedimiento almacenado sp_Carga_Bases_Consumos creará un índice a la tabla BaseConsumoPre luego de la importación para disminuir el tiempo del proceso de importación. El índice permite encontrar en menor tiempo los registros a eliminar. Se eliminará los registros que tengan fecha de última detección fuera del rango establecido. Se eliminará los registros que sean repetidos y que tengan el mayor sector.

En el cuadro, se detalla el procedimiento.

Tabla 30: Descripción de Procedimiento
SP_0_CARGA_BASES_CONSUMOS

NOMBRE STORE PROCEDURE	sp_0_Carga_Bases_Consumos	
PARAMETROS DECLARADOS	TIPO	DESCRIPCION
@NumeroCuenta	nvarchar(255)	Este parámetro guarda los números de cuenta repetidos.
@sector	nvarchar(255)	Este parámetro guarda el sector máximo de cada número de cuenta repetidos.
@total	Int	Este parámetro es guarda el total de Números de cuenta repetidos.
@inicial	Int	Es un contador que va recorriendo la tabla de número de cuentas repetidas.
@AnnoMinimo	Int	Es el año mínimo que puede tener la Fecha de ultima detección
@AnnoMaximo	Int	Es el año máximo que puede tener la Fecha de ultima detección
@fechaMin	datetime	Es la fecha de primera detección de la tabla periodo actual para el proceso de entrenamiento
@fechaMax	datetime	Es la fecha de fin de la tabla periodo actual para el proceso de entrenamiento
@contador	Int	Este parámetro mostrara las filas que se han eliminando.

SUB PROCEDURE	STORE	PARAMETROS RECIBE	PARAMETROS DEVUELVE	DESCRIPCION
TABLAS ACTUALIZADAS		Campos	SENTENCIAS	DESCRIPCION
BaseConsumoPre		[Fecha_Ult_Deteccion]	Delete	Se elimina los registros que tengan Fecha_Ult_Deteccion diferentes de nulo y sean mayores al año maximo y menores que año minimo.
		Sector	Delete	Elimina los registros que sean repetidos y que tengan el mayor sector
BaseConsumo		Todos	Insert	Inserta los datos convertidos según la estructura de la tabla BaseConsumo

Elaboración: La autora

Procedimiento sp_0_Carga_Valorizacion

El procedimiento almacenado sp_Carga_Valorizacion hace uso de la tabla ValorizacionPre (previamente importada) y se encargar de eliminar los registros que se encuentran fuera del rango establecido y de transferir los datos a la estructura establecida en la tabla Valorización.

En la tabla, se detalla el procedimiento de carga valorización.

Tabla 31: Descripción de Procedimiento SP_0_CARGA_VALORIZACION

NOMBRE STORE PROCEDURE	sp_0_Carga_Valorizacion		
PARAMETROS DECLARADOS	TIPO	DESCRIPCION	
@NumeroCuenta	nvarchar(255)	Este parámetro guarda los números de cuenta repetidos.	
@contador	Int	Es un contador que va almacenando el número de registros que se van a eliminar de la tabla ValorizacionPre en el proceso de limpieza de la data.	
@num	Int	Este parámetro es guarda el total de registros con fechas fuera de rango.	
@p	Int	Es un contador que va recorriendo la tabla que contiene los registros de fechas fuera de rango.	
@fechaMin	datetime	Es la fecha de primera detección de la tabla periodo actual para el proceso de entrenamiento	
@fechaMax	datetime	Es la fecha de fin de la tabla periodo actual para el proceso de entrenamiento	
@contador	Int	Este parámetro mostrara las filas que se han eliminado.	
TABLAS ACTUALIZADAS	Campos	SENTENCIA S	DESCRIPCION
ValorizaciónPre	[NumeroCuenta]	Delete	Se elimina los registros que cuyos número de cuenta correspondan a fechas fuera de rango.
Valorización	Todos	Insert	Inserta los datos convertidos según la estructura de la tabla Valorización.

Elaboración: La autora

Procedimiento sp_1_preparar_datos

Este procedimiento se inicia con la cargar de la tabla BasesConsumos a la tabla Consumo y ha esta última se le realiza una serie de pasos de preparación de datos.

Pasos de preparación:

1. Se carga la columna código facturación como la unión de todas las claves de facturación del cliente. Ejemplo de un código de facturación (NN***TTNNUN*TN).
2. Se reemplaza dentro del código de facturación las claves de facturación (T) por (N), debido a que estas claves (T) indican que fueron calculadas teóricamente o manualmente tal como se leyó en el momento. Del ejemplo anterior debería quedar así (NN***NNNNUN*NN).
3. Se eliminan los códigos de facturación que pertenezcan a clientes que tengan menos de doce meses (K=12) de habilitados debido a que este cliente no tendría historia que analizar.
4. Se actualiza el campo promedio consumo mensual para todos los registros. Donde promedio consumo de un cliente es la sumatoria de todos sus consumos entre la cantidad de meses.
5. Se eliminan los registros que contengan todos los días de facturación igual a cero.
6. Se eliminan los registros que contengan dentro del código de facturación una clave de facturación cerrada (U).
7. Se eliminan los registros que contengan un código de facturación que termine en (*), debido a que no se podría calcular su consumo promedio ya que se necesita saber el siguiente mes para calcular el consumo real.
8. Se eliminarán los registros que tengan el promedio consumo mensual menor que el consumo mínimo permitido establecido como parámetro inicialmente.
9. Se calculará el consumo promediado mes para los registros que contengan dentro del código de facturación un (*) seguido por una (N).
Analizando consumo promediado mes.
10. Se calcula los días de facturación real restando la fecha de lectura del actual menos la fecha de lectura del mes siguiente. Debido a que al hacer un análisis de este campo se detectó días de facturación desde 28 hasta 240 días para un mes. Ejemplo para Día Fact₂₃=Fec Lec₂₃-Fec Lec₂₄.

11. Se eliminan los días Fac_24 que sean mayores que el parámetro establecido inicialmente (días máximo=33), debido a que al hacer un histograma a los campos días de facturación arrojó valores entre 28 y 33 días el cual se corroboró con el cliente y en efecto era el intervalo que se trabajaba, y por ser el último mes de análisis no se tiene los días de facturación del siguiente mes para calcular el día real para el mes 24.
12. Se eliminan los registros que tengan algún día de facturación igual a cero, debido a que cuando se realice el cálculo de consumo promedio diario (Valor Consumo/ días de facturación) se produciría un resultado indeterminado.

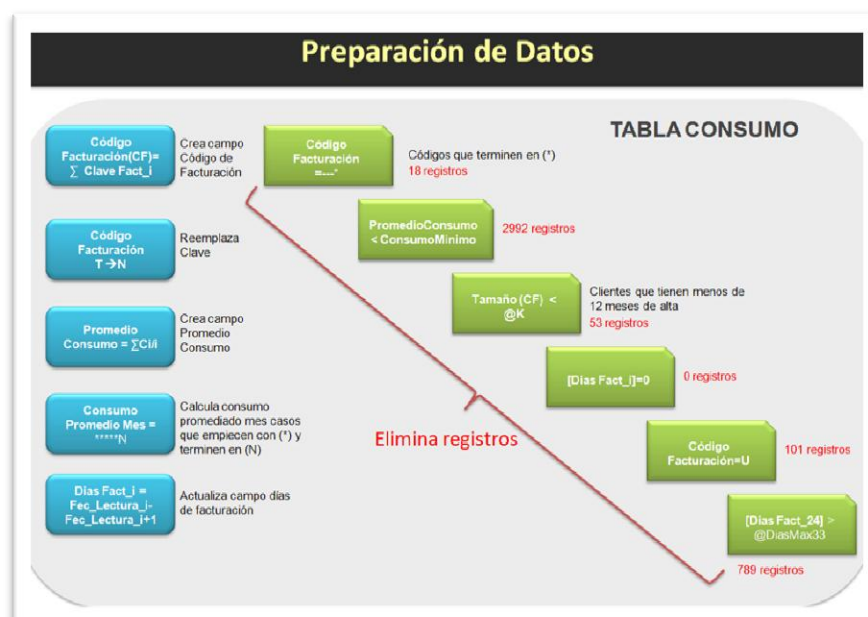


Figura 41: Proceso de Preparación de Datos.

Elaboración: La autora

A continuación, se presenta el cuadro donde se describe al detalle el procedimiento de preparación de datos.

Tabla 32: Descripción de procedimiento Sp_1_Prepara_Datos

NOMBRE STORE PROCEDURE		sp_1_prepara_datos		
PARAMETROS QUE RECIBE	TIPO	DESCRIPCIÓN		
@ConsumoMinimo	decimal(20,4)			
@K	int			
@DiasMax33	int			
PARAMETROS DECLARADOS	TIPO	DESCRIPCIÓN		
@NumeroCuenta	int	Este parámetro guarda los números de cuenta de los registros que se encuentran en la tabla Consumo		
@CFbase	varchar(255)	Este parámetro guarda el código de facturación inicial		
@CF	varchar(255)	Este parámetro guarda el código de facturación		
@Base	Int	Guarda la posición inicial del @CFbase		
@Rango	varchar(255)	Guarda el tamaño del @CFbase		
@pIniCF	Int	Este parámetro guarda la posición inicial		
@pFinCF	Int	Este parámetro guarda la posición Final		
@ids	Int	Este parámetro guarda la cantidad de registros que tengan un * en el código de facturación y un * al final		
@id	Int	Este parámetro es un contador		
@lenCF	int	Este parámetro guarda el tamaño del código de facturación		
@CFant	varchar(255)	Este parámetro guarda un rango del código de facturación.		
SUB STORE PROCEDURE	PARAMETROS RECIBE	PARAMETROS DEVUELVE	DESCRIPCIÓN	
➤ p_1rango_codigo_facturacion		@CF	Este proceso devuelve un rango del código de facturación y sus posiciones inicial y final para poder ser utilizadas en el proceso de	
		@Base		
		@pIniCF		
		@pFinCF		

			Consumos promediados
➤ p_1prepara_consumo _promediado_mes_su b	@NumeroCuenta		Este proceso recibe los parámetros entregados por el proceso sp_1rango_codigo_facturacion y actualiza los Consumos Promediados
	@pIniCF		
	@pFinCF		
	@lenCF		
TABLAS ACTUALIZADAS	Campos	SENTENCIAS	DESCRIPCIÓN
Consumo	Todos	Insert	Carga los datos de la tabla BaseConsumo y los guarda en Consumo
	CodigoFacturacion	delete	Elimina los registros que no permiten la elaboración del Modelo ver Tabla Análisis del Proceso de eliminación
	PromedioConsumo		

Elaboración: La autora

Procedimiento sp_2_Carga_Hurto_Maestro

Este proceso carga los consumos promedios diarios de los clientes de la tabla Consumo a la tabla Maestro con estado habilitado y tarifa BT5 B. La carga se efectúa consultando en la tabla valorización los campos fecha de inicio y fin de recuperacion de caso de tratarse de un cliente hurtador.

Se actualiza los campos hurtold con el valor 'si', y en el campo Hurto_i, se coloca 'Si' durante todo el periodo de inicio y fin de recuperacion como se muestra en la figura 15 por otro lado se actualiza el campo HurtoMeses (cantidad de meses que el cliente hurtó) así por ejemplo considerando que un cliente tiene una valorización entre los días 12 de noviembre del 2007 y el 16 de abril del 2008 estos se traducen en la tabla Maestro como 'si' entre los meses 18 a 23.

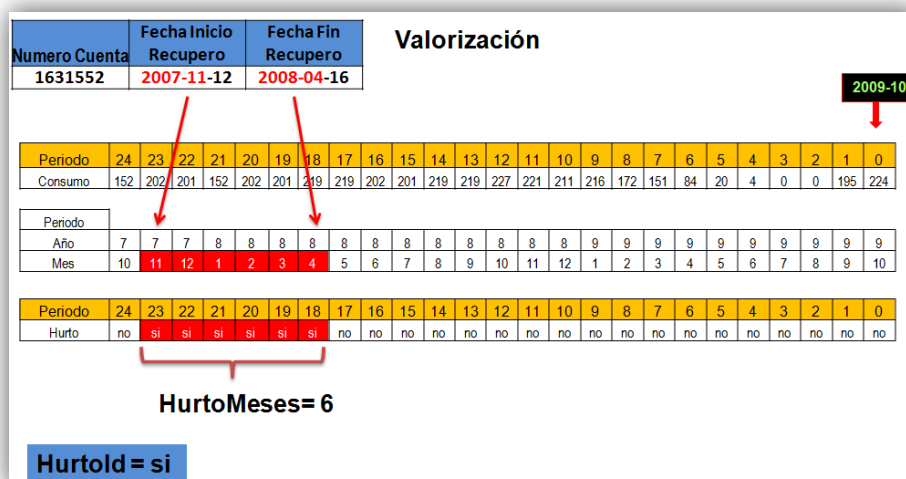


Figura 42: Proceso de Carga Hurtos a la tabla Maestro

Elaboración: La autora

A continuación, se muestra en el cuadro el detalle los parámetros y sub procesos que utiliza el procedimiento carga hurto maestro.

Tabla 33: Descripción de Procedimiento SP_2_Carga_Hurto_Maestro

NOMBRE PROCEDURE	STORE	sp_2_Carga_Hurto_Maestro		
PARAMETROS RECIBE	QUE	TIPO	DESCRIPCION	
@Fechalni		Datetime		
@FechaFin		Datetime		
PARAMETROS DECLARADOS		TIPO	DESCRIPCION	
@idv		Int	Este parámetro es un contador de los registros que se encuentran en la tabla valorización.	
@size		Int	Este parámetro guarda el total de registros que se encuentran en la tabla valorización.	
@NumeroCuenta		Int	Guarda los números de cuenta de la tabla valorización.	
SUB STORE PROCEDURE		PARAMETROS RECIBE	PARAMETROS DEVUELVE	DESCRIPCION
sp_2carga_hurto_maestro_sub		@NumeroCuenta		Este store procesa los parámetros recibidos y actualiza la tabla
		@Fechalni		
		@FechaFin		

			Maestro los campos Hurtold, Meses y hurtos por cada mes.
TABLAS ACTUALIZADAS	Campos	SENTENCIAS	DESCRIPCION
FiltroMaestro	Tarifa	insert	Ingresa datos a la tabla FiltroMaestro
	Valor		
	Estado		
Maestro	EstadoCliente Tarifa	Insert	Se filtra a los cliente habilitados y tarifa BT5 B
	Hurtold	Update	Se actualiza con Si
	Meses	Update	Se actualiza con la cantidad de meses que se hurto.
	Hurto_i	Update	Se actualiza mes a mes si hubo un hurto

Elaboración: La autora

Procedimiento sp_3_carga_hurto_casos

Este es el proceso principal del modelo donde los casos de hurto almacenados en la tabla intermedia **Maestro** se cargará a la tabla **Consulta** que contendrá todos los patrones de comportamiento de los consumos que se extraerán transacciones que consideran (K-1=11) meses de datos históricos y 1 mes (ultimo o más reciente para predecir).

Este procedimiento se encarga de cargar todos los casos de hurto de la tabla Maestro y analizar hacia atrás K-1 periodos (K es un parámetro asignado al inicio para saber el rango de análisis).

Tabla 34: Descripción de Procedimiento sp_3_carga_hurto_casos

NOMBRE STORE PROCEDURE	sp_3_carga_hurto_casos		
PARAMETROS QUE RECIBE	TIPO	DESCRIPCION	
@K	Int	Rango de tiempo a analizar para la crear el modelo.	
PARAMETROS DECLARADOS	TIPO	DESCRIPCION	
@contador	Int		
@id	Int	Este parámetro será un contador recorrerá registro por registro a los hurtadores.	
@NumeroCuenta	Int		
@ocurrencias	Int	Este parámetro guarda el numero total de hurtadores.	
@Hurto_i	varchar(255)	Este parámetro guarda los 24 hurtos.	
@i	Int		
@query	varchar(255)		
@querycliente	varchar(255)	Este parámetro guarda los datos NumeroCuenta, HurtoMeses,Hurtold	
@queryhurto	varchar(255)	Este parámetro guarda los 25 hurtos de cada cliente	
@queryconsumo	varchar(255)	Este parámetro guarda los 25 consumos de cada cliente	
@c	varchar(255)	Este parámetro guarda una comilla simple.	
SUB STORE PROCEDURE	PARAMETROS RECIBE	PARAMETROS DEVUELVE	DESCRIPCION
sp_3carga_hurto_Consulta_hurto_sub	@NumeroCuenta	@contador	Este sub proceso se encarga de colocar Hurtoldk con valor 'Si', si se detecto un hurto al último mes y envía un contador para hacer una análisis hacia atrás.
	Numero		
	@NumeroCuenta		

TABLAS ACTUALIZADAS	Campos	SENTENCIAS	DESCRIPCION
Consulta	Todos	Insert	Inserta los casos de hurto analizando si existe un hurto al último mes, guarda sus consumo de ese mes hacia atrás k periodos, luego analiza el inmediato anterior si existe hurto hace el mismo procedimiento luego inserta todos los casos que nunca tuvieron un hurto.

Elaboración: La autora

Tabla Consultada

periodo	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
año	7	7	7	7	8	8	8	8	8	8	8	8	8	8	8	8	9	9	9	9	9	9	9	9	9	9
mes	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10
consumo	120	58	130	164	127	28	220	222	173	6	70	144	248	56	25	207	252	48	178	156	269	151	120	54	298	44
hurto	si	si	no	no	no	no	no	no	no	no	no	si	no	no	no	si	si	si	no	no	no	no	no	no	no	no

periodo	11	10	9	8	7	6	5	4	3	2	1	0
consumo	222	173	6	70	144	248	56	25	207	252	48	178
hurto	no	no	no	no	no	si	no	no	no	si	si	si

220	222	173	6	70	144	248	56	25	207	252	48
no	no	no	no	no	no	si	no	no	no	si	si

28	220	222	173	6	70	144	248	56	25	207	252
no	no	no	no	no	no	si	no	no	no	no	si

58	130	164	127	28	220	222	173	6	70	144	248
si	no	no	no	no	no	no	no	no	no	no	si

K=12

Figura 43: Extraer Casos de Hurto

Elaboración: La autora

A continuación, se muestra el cuadro donde se detalla el proceso de carga hurtos, además de los casos de hurto, este procedimiento también carga los casos de no hurto desde la tabla Maestro, así como también los parámetros y los sub procesos que utiliza.

Fase 4: Modelado

En esta fase, varias técnicas de modelado son aplicadas, calibradas.

Típicamente hay varias técnicas para el mismo tipo de problema de minería

de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Esta fase se relaciona con la fase de evaluación, todo modelo debe ser evaluado para calcular su rendimiento.

Descripción de Modelado

Para crear el modelo de detección de fraudes de energía eléctrica en clientes residenciales para el aprendizaje de los comportamientos se utilizó la Red Neuronal de Perceptrón Multicapa – Retro propagación (PMC).

La estructura de PMC está dividida por capas las cuales son:

- Capa de entrada
- Capa oculta
- Capa de salida

En este tipo de red neuronal artificial, se ingresa un número de entradas que van conforme al número de neuronas que hay en la capa de entrada, después estas se conectan con las neuronas de la capa oculta por medio de vértices, en donde cada salida de las neuronas, presentes en la capa de entrada, se asocia a cada una de las neuronas en la capa oculta.

Una vez que están conectadas todas las salidas de la capa de entrada a las entradas de las neuronas de la capa oculta se repite el mismo procedimiento, en caso de que se presenten más capas ocultas; si no es el caso las salidas de cada neurona en la capa oculta es conectada a la capa de salida, de la misma forma que se describió, y finalmente las salidas finales del PMC. En la Figura 43 se muestra el esquema de la PMC.

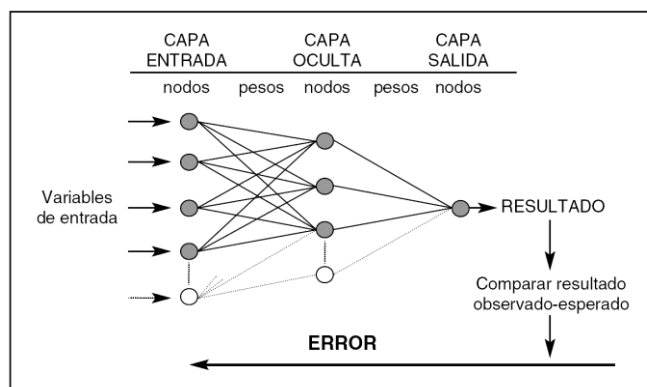


Figura 44: Modelo del Perceptrón Multicapa - Retropropagación.

Fuente: Moreno B. (2010). Minería de Datos en grandes bases de datos. (p74)

Funcionamiento del Perceptron Multicapa – Retropropagación

Según Moreno B.(2009), explicó que “El término retropropagación se basa en el método del gradiente descendiente para encontrar el error en una red hacia adelante (feed-foward, de aprendizaje supervisado, en donde se necesita un conjunto de entrenamiento y el valor de salida esperada)”.

El funcionamiento de este tipo de redes neuronales artificiales se puede dividir en las siguientes:

- Los datos de entrenamiento se pasan hacia adelante (forward pass), las salidas son evaluadas calculando el error en cada caso.
- Se realiza entonces el paso hacia atrás (backward pass), en donde el error calculado en la capa de salida, se utiliza para cambiar el peso de cada capa ocultas de la red neuronal, hasta llegar a la capa de salida, evaluando recursivamente los gradientes locales para cada neurona.

Al final de estas dos etapas, se tiene un PMC entrenado.

Modelo de red neuronal multicapa

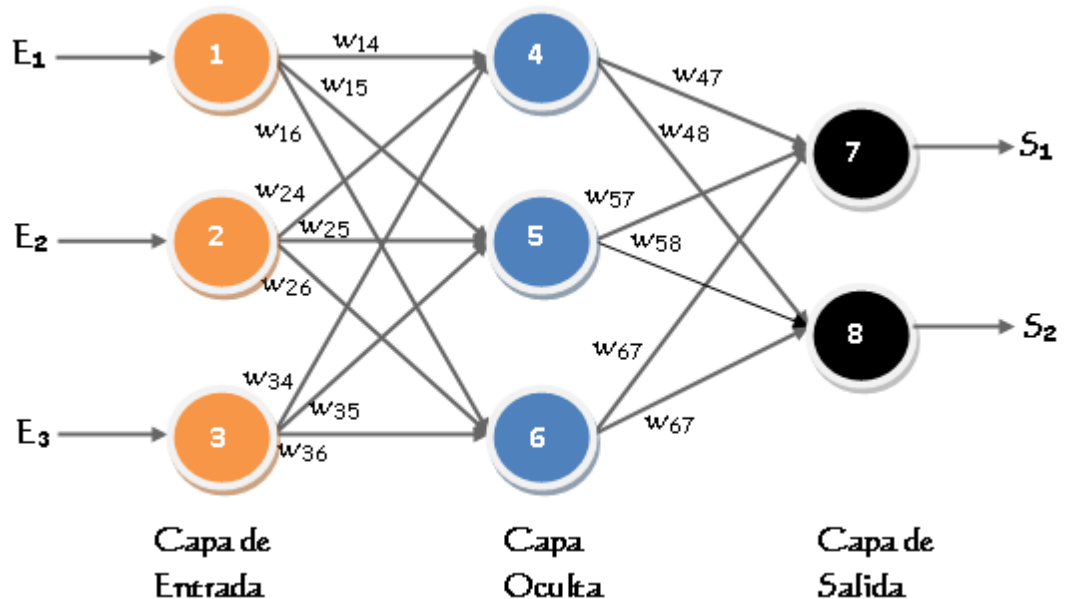


Figura 45: Modelo de red neuronal Multicapa.

Fuente: Kasperu(2010)

Fórmulas para crear el modelo de red neuronal:

A. Para calcular las salidas de cada neurona hacia adelante.

Ejemplo:

Para calcular salidas de las neuronas de cada capa

Por ejemplo para el nodo 4

$$\text{Nodo 4} = \sum (E1 \cdot W14 + E2 \cdot W24 + E3 \cdot W34) = SO4$$

Donde:

- E1= Entrada 1
- W14= Peso del entrada 1 al nodo 4
- SO = Salida Obtenida

Calculo de salida o resultado del nodo 4 con la función de transferencia

$$f(x) = \frac{1}{1 + e^{-x}}$$

$$F(O4) = 1 / (1 + e^{-(O4)}) = RO4$$

Donde:

RO= Resultado obtenido

Luego una vez que se tiene el resultado obtenido de la capa intermedia, estos valores corresponden a los valores de entrada de la capa de salida y se continúa calculando el resultado obtenido del nodo 7 y 8.

Una vez obtenido el resultado de los nodos 7 y 8, se compara con los valores esperados y se empieza el proceso de retropropagación hacia atrás calculando los errores y los valores de g_i , la capa de salida y después ir a cada una de las capas de la red.

Fórmula para g_q unidades en la capa de salida.

$$\delta_q = (\delta_q - O_q) O_q (1 - O_q)$$

Ejemplo: Para el nodo 6 tenemos.

$$\delta_6 = (\delta_6 - O_6) O_6 (1 - O_6)$$

Para cada una de las q's unidades en la capa oculta, retropropagamos los valores de las g:

$$\delta_q = O_q(1 - O_q) \sum_i W_{iq} \delta_i$$

Para el nodo 3 tenemos:

$$\delta_3 = O_3(1 - O_3) [w_{63} \delta_6 + w_{73} \delta_7]$$

Ahora lo que debemos de hacer es actualizar los pesos de todas las capas con la siguiente fórmula:

$$W_{qp}^{nuevo} = W_{qp}^{viejo} + \Delta W_{qp}$$

Nuestra $\eta = 0.1$, por lo que tendríamos para cada uno de los pesos:

Por ejemplo:

Para el peso W_{63} tenemos:

$$W_{63}^{nuevo} = W_{63}^{viejo} + \Delta W_{63} = W_{63}^{viejo} + \eta \delta_6 O_3$$

Finalmente, los pesos para la capa oculta se actualizan, esto se realiza para la siguiente capa hasta finalizar obteniendo un PMC entrenado con las entradas.

Datos utilizados para entrenar los modelos de Red Neuronal Multicapa

Las variables independientes y dependientes, para entrenar el model, deben tener la estructura en columnas

Datos de entrada:

- Número de cuenta, identificados del cliente. (No entra a la red, solo es un identificador interno)
- 12 consumos del cliente.

Dato de salida

- HurtoldK

Numero Cuenta	C_11	C_10	C_9	C_8	C_7	C_6	C_5	C_4	C_3	C_2	C_1	C_0	Hurto
3	7.96	8.43	8.12	7.31	7.07	7.16	8.42	4.79	6.59	6.56	7.03	0.00	no
4	26.58	28.48	22.88	30.73	17.40	17.86	17.22	13.53	12.34	11.16	15.32	13.23	no
5	5.50	5.93	5.58	7.48	7.34	6.58	6.00	5.93	6.17	5.88	5.87	5.76	no
7	8.54	9.37	8.12	7.93	8.28	8.31	8.00	7.72	6.38	6.41	6.27	5.81	no
825269	2.66	2.42	2.39	2.36	0.00	0.00	0.58	0.00	2.40	0.00	0.95	0.95	si
866531	7.20	6.41	7.00	7.03	7.00	6.62	6.78	6.36	5.93	4.32	2.94	0.52	si
191544	3.35	3.57	3.83	3.41	2.81	2.94	3.10	2.93	2.68	3.59	3.84	3.55	si
628145	8.35	9.80	11.41	10.33	4.88	6.73	9.86	10.39	8.94	7.77	8.34	0.90	si

Figura 46: Estructura de Datos para Entrenar Modelo

Elaboración: La autora

Modelo Desarrollado

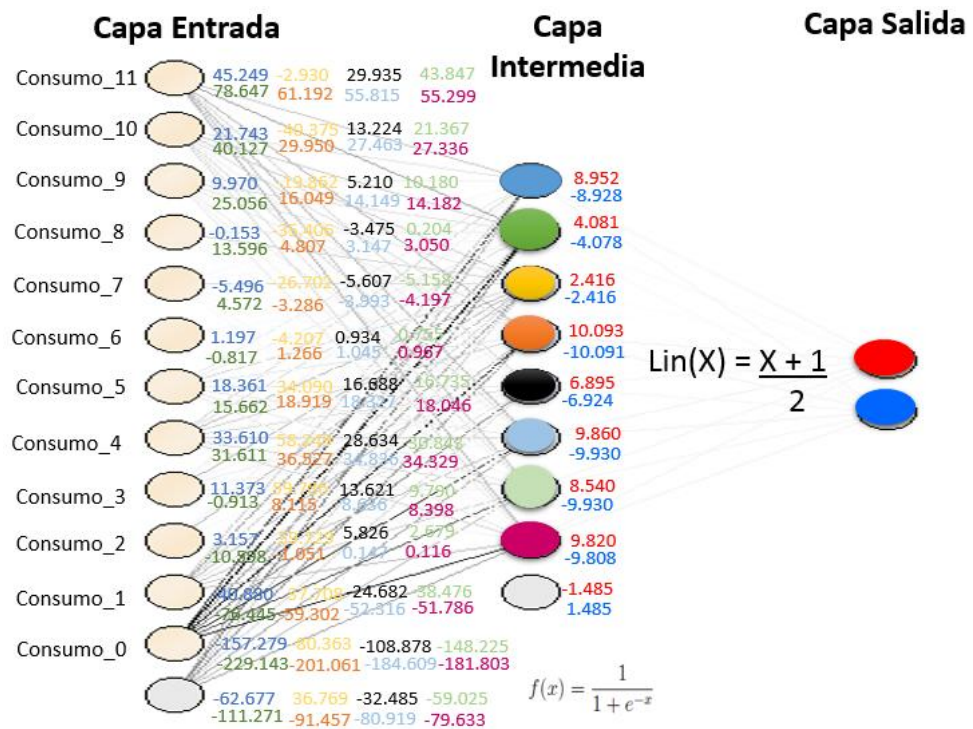


Figura 47: Modelo de Red Neuronal

Elaboración: La autora

Cálculo de Pesos por cada Nodo

Con la herramienta de Análisis Services se creó el modelo y con la herramienta Rapidminer se extrajeron los pesos del modelo.

ImprovedNeuralNet

Hidden 1
=====

Node 1 (Sigmoid)

Consumo_11: 45.249
Consumo_10: 21.743
Consumo_9: 9.970
Consumo_8: -0.153
Consumo_7: -5.496
Consumo_6: 1.197
Consumo_5: 18.361
Consumo_4: 33.610
Consumo_3: 11.373
Consumo_2: 3.157
Consumo_1: -40.880
Consumo_0: -157.279
Threshold: -62.677

Node 2 (Sigmoid)

Consumo_11: 78.647
Consumo_10: 40.127
Consumo_9: 25.056
Consumo_8: 13.596
Consumo_7: 4.572
Consumo_6: -0.817
Consumo_5: 15.662
Consumo_4: 31.611
Consumo_3: -0.913
Consumo_2: -10.598
Consumo_1: -76.445
Consumo_0: -229.143
Threshold: -111.271

Node 3 (Sigmoid)

Consumo_11: -2.930
Consumo_10: -40.375
Consumo_9: -19.862
Consumo_8: -35.406
Consumo_7: -26.702
Consumo_6: -4.207
Consumo_5: 34.090
Consumo_4: 58.249
Consumo_3: 59.786
Consumo_2: 59.729
Consumo_1: 37.708
Consumo_0: -80.363
Threshold: 36.769

Node 4 (Sigmoid)

Consumo_11: 61.192
Consumo_10: 29.950
Consumo_9: 16.049
Consumo_8: 4.807
Consumo_7: -3.286
Consumo_6: 1.266
Consumo_5: 18.919
Consumo_4: 36.527
Consumo_3: 8.115
Consumo_2: -1.051
Consumo_1: -59.302
Consumo_0: -201.061
Threshold: -91.457

Node 5 (Sigmoid)

Consumo_11: 29.935
Consumo_10: 13.224
Consumo_9: 5.210
Consumo_8: -3.475
Consumo_7: -5.607
Consumo_6: 0.934
Consumo_5: 16.688
Consumo_4: 28.634
Consumo_3: 13.621
Consumo_2: 5.826
Consumo_1: -24.682
Consumo_0: -108.878
Threshold: -32.485

Node 6 (Sigmoid)

Consumo_11: 55.815
Consumo_10: 27.463
Consumo_9: 14.149
Consumo_8: 3.147
Consumo_7: -3.993
Consumo_6: 1.045
Consumo_5: 18.327
Consumo_4: 34.836
Consumo_3: 8.636
Consumo_2: 0.147
Consumo_1: -52.316
Consumo_0: -184.609
Threshold: -80.919

Node 7 (Sigmoid)

Consumo_11: 43.847
Consumo_10: 21.367
Consumo_9: 10.180
Consumo_8: 0.204
Consumo_7: -5.158
Consumo_6: 0.755
Consumo_5: 16.735
Consumo_4: 30.843
Consumo_3: 9.790
Consumo_2: 2.679
Consumo_1: -38.476
Consumo_0: -148.225

Threshold: -59.025

Node 8 (Sigmoid)

Consumo_11: 55.299
Consumo_10: 27.336
Consumo_9: 14.182
Consumo_8: 3.050
Consumo_7: -4.197
Consumo_6: 0.967
Consumo_5: 18.046
Consumo_4: 34.329
Consumo_3: 8.398
Consumo_2: 0.116
Consumo_1: -51.786
Consumo_0: -181.803
Threshold: -79.633

Output

=====

Class 'si' (Sigmoid)

Node 1: 8.952
Node 2: 4.081
Node 3: 2.416
Node 4: 10.093
Node 5: 6.895
Node 6: 9.860
Node 7: 8.540
Node 8: 9.820
Threshold: -1.485

Class 'no' (Sigmoid)

Node 1: -8.928
Node 2: -4.078
Node 3: -2.416
Node 4: -10.091
Node 5: -6.924
Node 6: -9.930
Node 7: -9.930
Node 8: -9.808
Threshold: 1.485

Fórmula del Modelo Predictivo:

$Y(xSI) = f_{lineal} ($

$f_{sigmoidClaseSI} (\sum ($

$f_{sigmoidN1} (\sum (Consumo_{11} * 45.249 + Consumo_{10} * 21.743 + Consumo_{9} * 9.970 + Consumo_{8} * -0.153 + Consumo_{7} * -5.496 + Consumo_{6} * 1.197 + Consumo_{5} * 18.361 + Consumo_{4} * 33.610 + Consumo_{3} * 11.373 + Consumo_{2} * 3.157 + Consumo_{1} * -40.880 + Consumo_{0} * -157.279 - 62.677) \text{ Nodo } 1) * (8.952) +$

$$f_{\text{sigmoidN2}}(\sum (\text{Consumo}_{11} * 78.647 + \text{Consumo}_{10} * 40.127 + \text{Consumo}_9 * 25.056 + \text{Consumo}_8 * 13.596 + \text{Consumo}_7 * 4.572 + \text{Consumo}_6 * -0.817 + \text{Consumo}_5 * 15.662 + \text{Consumo}_4 * 31.611 + \text{Consumo}_3 * -0.913 + \text{Consumo}_2 * -10.598 + \text{Consumo}_1 * -76.445 + \text{Consumo}_0 * -229.143 - 111.271) \text{Nodo } 2) * (4.081) +$$

$$f_{\text{sigmoidN3}}(\sum (\text{Consumo}_{11} * -2.930 + \text{Consumo}_{10} * -40.375 + \text{Consumo}_9 * -19.862 + \text{Consumo}_8 * -35.406 + \text{Consumo}_7 * -26.702 + \text{Consumo}_6 * -4.207 + \text{Consumo}_5 * 34.090 + \text{Consumo}_4 * 58.249 + \text{Consumo}_3 * 59.786 + \text{Consumo}_2 * 59.729 + \text{Consumo}_1 * 37.708 + \text{Consumo}_0 * -80.363 - 36.769) \text{Nodo } 3) * (2.416)$$

$$f_{\text{sigmoidN4}}(\sum (\text{Consumo}_{11} * 61.192 + \text{Consumo}_{10} * 29.950 + \text{Consumo}_9 * 16.049 + \text{Consumo}_8 * 4.807 + \text{Consumo}_7 * -3.286 + \text{Consumo}_6 * 1.266 + \text{Consumo}_5 * 18.919 + \text{Consumo}_4 * 36.527 + \text{Consumo}_3 * 8.115 + \text{Consumo}_2 * -1.051 + \text{Consumo}_1 * -59.302 + \text{Consumo}_0 * -201.061 - 91.457) \text{Nodo } 4) * (10.093)$$

$$f_{\text{sigmoidN5}}(\sum (\text{Consumo}_{11} * 29.935 + \text{Consumo}_{10} * 13.224 + \text{Consumo}_9 * 5.210 + \text{Consumo}_8 * -3.475 + \text{Consumo}_7 * -5.607 + \text{Consumo}_6 * 0.934 + \text{Consumo}_5 * 16.688 + \text{Consumo}_4 * 28.634 + \text{Consumo}_3 * 13.621 + \text{Consumo}_2 * 5.826 + \text{Consumo}_1 * -24.682 + \text{Consumo}_0 * -108.878 - 32.485) \text{Nodo } 5) * (6.895)$$

$$f_{\text{sigmoidN6}}(\sum (\text{Consumo}_{11} * 29.935 + \text{Consumo}_{10} * 13.224 + \text{Consumo}_9 * 5.210 + \text{Consumo}_8 * -3.475 + \text{Consumo}_7 * -5.607 + \text{Consumo}_6 * 0.934 + \text{Consumo}_5 * 16.688 + \text{Consumo}_4 * 28.634 + \text{Consumo}_3 * 13.621 + \text{Consumo}_2 * 5.826 + \text{Consumo}_1 * -24.682 + \text{Consumo}_0 * -108.878 - 32.485) \text{Nodo } 6) * (9.860)$$

$$f_{\text{sigmoidN7}}(\sum (\text{Consumo}_{11} * 43.847 + \text{Consumo}_{10} * 21.367 + \text{Consumo}_9 * 10.180 + \text{Consumo}_8 * 0.204 + \text{Consumo}_7 * -5.158 + \text{Consumo}_6 * 0.755 + \text{Consumo}_5 * 16.735 + \text{Consumo}_4 * 30.843 + \text{Consumo}_3 * 9.790 + \text{Consumo}_2 * 2.679 + \text{Consumo}_1 * -38.476 + \text{Consumo}_0 * -148.225 - 59.025) \text{Nodo } 7) * (8.540)$$

$$f_{\text{sigmoidN8}}(\sum (\text{Consumo}_{11} * 55.299 + \text{Consumo}_{10} * 27.336 + \text{Consumo}_9 * 14.182 + \text{Consumo}_8 * 3.050 + \text{Consumo}_7 * -4.197 + \text{Consumo}_6 * 0.967 + \text{Consumo}_5 * 18.046 + \text{Consumo}_4 * 34.329 + \text{Consumo}_3 * 8.398 + \text{Consumo}_2 * 0.116 + \text{Consumo}_1 * -51.786 + \text{Consumo}_0 * -181.803 - 79.633) \text{Nodo } 4) * (9.820) - 1.485$$

)

$$Y(xNO) = f_{lineal} ($$

$$f_{\text{sigmoidClaseNO}} (\sum ($$

$$f_{\text{sigmoidN1}} (\sum (\text{Consumo}_{11} * 45.249 + \text{Consumo}_{10} * 21.743 + \text{Consumo}_9 * 9.970 + \text{Consumo}_8 * -0.153 + \text{Consumo}_7 * -5.496 + \text{Consumo}_6 * 1.197 + \text{Consumo}_5 * 18.361 + \text{Consumo}_4 * 33.610 + \text{Consumo}_3 * 11.373 + \text{Consumo}_2 * 3.157 + \text{Consumo}_1 * -40.880 + \text{Consumo}_0 * -157.279 - 62.677) \text{Nodo 1}) * (-8.928) +$$

$$f_{\text{sigmoidN2}} (\sum (\text{Consumo}_{11} * 78.647 + \text{Consumo}_{10} * 40.127 + \text{Consumo}_9 * 25.056 + \text{Consumo}_8 * 13.596 + \text{Consumo}_7 * 4.572 + \text{Consumo}_6 * -0.817 + \text{Consumo}_5 * 15.662 + \text{Consumo}_4 * 31.611 + \text{Consumo}_3 * -0.913 + \text{Consumo}_2 * -10.598 + \text{Consumo}_1 * -76.445 + \text{Consumo}_0 * -229.143 - 111.271) \text{Nodo 2}) * (-4.078) +$$

$$f_{\text{sigmoidN3}} (\sum (\text{Consumo}_{11} * -2.930 + \text{Consumo}_{10} * -40.375 + \text{Consumo}_9 * -19.862 + \text{Consumo}_8 * -35.406 + \text{Consumo}_7 * -26.702 + \text{Consumo}_6 * -4.207 + \text{Consumo}_5 * 34.090 + \text{Consumo}_4 * 58.249 + \text{Consumo}_3 * 59.786 + \text{Consumo}_2 * 59.729 + \text{Consumo}_1 * 37.708 + \text{Consumo}_0 * -80.363 - 36.769) \text{Nodo 3}) * (-2.416)$$

$$f_{\text{sigmoidN4}} (\sum (\text{Consumo}_{11} * 61.192 + \text{Consumo}_{10} * 29.950 + \text{Consumo}_9 * 16.049 + \text{Consumo}_8 * 4.807 + \text{Consumo}_7 * -3.286 + \text{Consumo}_6 * 1.266 + \text{Consumo}_5 * 18.919 + \text{Consumo}_4 * 36.527 + \text{Consumo}_3 * 8.115 + \text{Consumo}_2 * -1.051 + \text{Consumo}_1 * -59.302 + \text{Consumo}_0 * -201.061 - 91.457) \text{Nodo 4}) * (-10.091)$$

$$f_{\text{sigmoidN5}} (\sum (\text{Consumo}_{11} * 29.935 + \text{Consumo}_{10} * 13.224 + \text{Consumo}_9 * 5.210 + \text{Consumo}_8 * -3.475 + \text{Consumo}_7 * -5.607 + \text{Consumo}_6 * 0.934 + \text{Consumo}_5 * 16.688 + \text{Consumo}_4 * 28.634 + \text{Consumo}_3 * 13.621 + \text{Consumo}_2 * 5.826 + \text{Consumo}_1 * -24.682 + \text{Consumo}_0 * -108.878 - 32.485) \text{Nodo 5}) * (-6.924)$$

$$f_{\text{sigmoidN6}} (\sum (\text{Consumo}_{11} * 29.935 + \text{Consumo}_{10} * 13.224 + \text{Consumo}_9 * 5.210 + \text{Consumo}_8 * -3.475 + \text{Consumo}_7 * -5.607 + \text{Consumo}_6 * 0.934 + \text{Consumo}_5 * 16.688 + \text{Consumo}_4 * 28.634 + \text{Consumo}_3 * 13.621 + \text{Consumo}_2 * 5.826 + \text{Consumo}_1 * -24.682 + \text{Consumo}_0 * -108.878 - 32.485) \text{Nodo 6}) * (-9.930)$$

$$f_{\text{sigmoidN7}} (\sum (\text{Consumo}_{11} * 43.847 + \text{Consumo}_{10} * 21.367 + \text{Consumo}_9 * 10.180 + \text{Consumo}_8 * 0.204 + \text{Consumo}_7 * -5.158 + \text{Consumo}_6 * 0.755 + \text{Consumo}_5 * 16.735 + \text{Consumo}_4 * 30.843 + \text{Consumo}_3 * 9.790 + \text{Consumo}_2 * 2.679 + \text{Consumo}_1 * -38.476 +$$

Consumo_0* -148.225
- 59.025) Nodo 7) * (-8.488)

$f_{\text{sigmoidN8}}(\sum (\text{Consumo}_{11} * 55.299 + \text{Consumo}_{10} * 27.336 + \text{Consumo}_9 * 14.182 + \text{Consumo}_8 * 3.050 + \text{Consumo}_7 * -4.197 + \text{Consumo}_6 * 0.967 + \text{Consumo}_5 * 18.046 + \text{Consumo}_4 * 34.329 + \text{Consumo}_3 * 8.398 + \text{Consumo}_2 * 0.116 + \text{Consumo}_1 * -51.786 + \text{Consumo}_0 * -181.803 - 79.633) \text{Nodo 4}) * (-9.808) + 1.485))$

Donde:

Función Sigmoidal:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Función Lineal:

$$\text{Lin}(X) = \frac{X + 1}{2}$$

Descripción de datos para entrenar y probar los modelos

1. Datos de consumos anuales para entrenar el modelo

Para entrenar el modelo, se extrajeron 30000 registros aleatorios de los cuales el 33% corresponde a clientes con casos de hurto y el 66% de clientes que nunca hurtaron (sin código de irregularidad). Estos porcentajes fueron el resultado de un análisis previo realizado con diversas proporciones de registros, siendo estas cantidades la que arrojaron mayor grado de rendimiento (su modificación debe ser reportada al proveedor porque afectaría el desempeño del Modelo).

2. Reporte de consumos anuales para testear el modelo

Para testear el modelo, se extrajeron 100000 registros entre hurtadores y no hurtadores del mismo periodo de los casos que se creó el modelo.

Para conocer el porcentaje de asertividad del Modelo se crea una Matriz de Clasificación, con la herramienta de complemento de Minería de Datos de Excel, se elige el modelo creado, columnas de datos a evaluar y columna a predecir.

Creación de origen de datos para crear Modelos

El origen de datos tiene por finalidad vincularse a una Base de datos, y acceder a las tablas y vistas que posee.

Es importante antes de crear un origen de datos, crear una base de datos en Analysis service que almacenara los modelos que se crearan en Excel y comprobar que exista conexión a esta base de datos.

Para crear la base de datos ingresamos a Microsoft Management Studio y elegimos *Analysis service* , luego colocamos sobre Databases y creamos una nueva base de datos de nombre, en este caso DMAddinsDB y finalizamos.

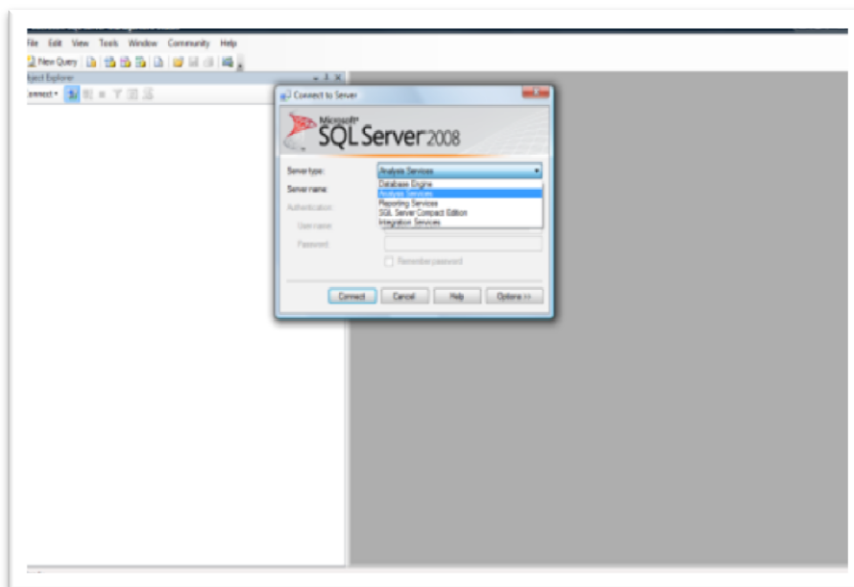


Figura 48: Conexión a Análisis Services.

Elaboración: La autora

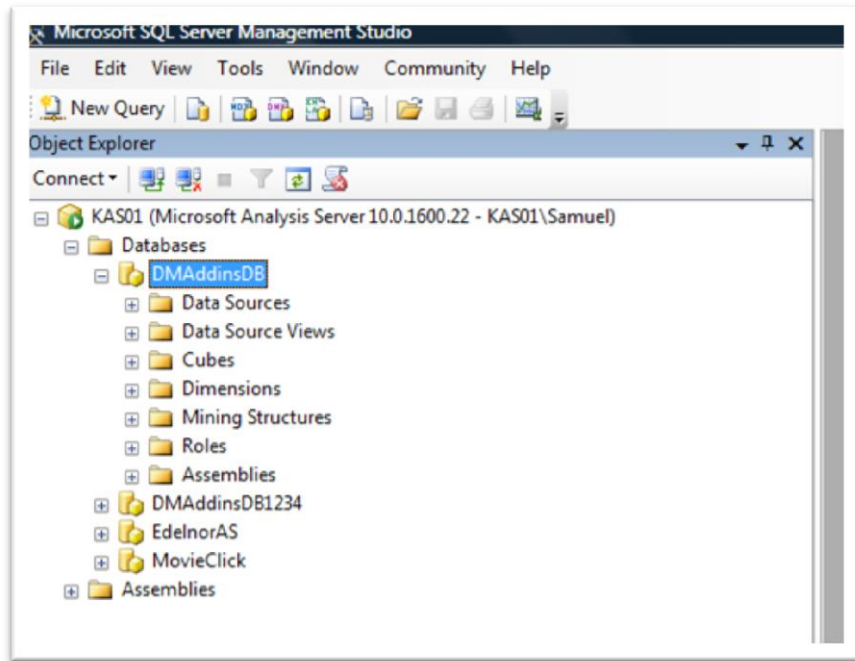


Figura 49: Creación de Base de Datos en Análisis Services.

Elaboración: La autora

Luego de creada la base de datos probamos que exista conexión con la BD creada en Análisis Service, vamos a Inicio → Todos los programas - → Complemento de minería de datos de Microsoft SQL a través del complemento de Minería de datos de SQL Server 2008 → Utilidad de configuración del servidor.

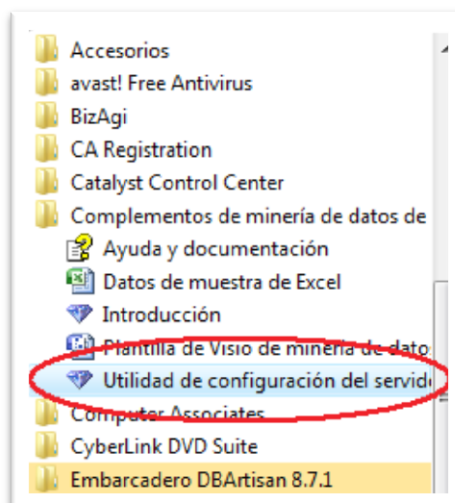


Figura 50: Probar Conexión con a la BD con el Complemento de Minería de Datos.

Elaboración: La autora

Nos mostrará la ventana de Configuración de Complementos de MD de SQL server 2008 para office, seleccionamos el nombre de nuestro servidor donde se instaló el sql server 2008., luego nos mostrará la opción de crear una base de datos, elegimos usar base de datos existente y seleccionamos la base de datos que creamos y finalizamos.

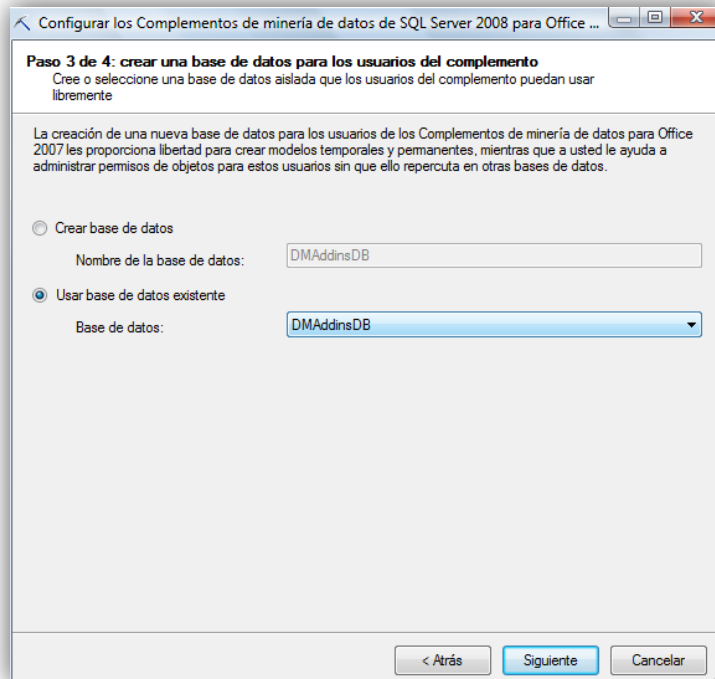


Figura 51: Conexión a la BD de Análisis Services.

Elaboración: La autora

Para crear un origen de datos en caso la información se encuentre en una Base de datos (tablas, Vistas), inicialmente los modelos se crearon con vistas a la base datos, para ello ingresamos a Microsoft Office Excel 2007 debemos tener habilitado la pestaña de **Minería de datos**. Se ingresa a la opción Clasificar.

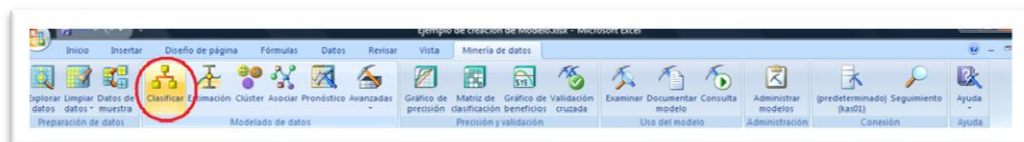


Figura 52: Creación de Origen de Datos - Clasificar Datos.

Elaboración: La autora

Se muestra el asistente para clasificar se elige **“Origen de datos Externos”** y nos mostrará un editor de consulta de origen de datos, creamos un nuevo origen de datos dando clic en el círculo rojo como muestra la figura de la izquierda. Luego nos muestra la ventana Nuevo Origen de datos del servidor de Analysis Services, ingresamos los datos en los círculos rojos como nos muestra la figura de la derecha.

- Nombre de Origen de datos: en este caso, la información que extraigamos de una Base Datos le colocamos BD_ModeloEdelnor, de preferencia un nombre que vaya relacionado a la base de datos que se desee vincular.
- Nombre del servidor: es el nombre del servidor donde se instaló el Microsoft SQL server 2008, en este caso el nombre es KAS01
- Credenciales de inicio de sesión: si para ingresar al servidor se usa alguna usuario con contraseña se debe seleccionar la opción Utilizar autenticación de SQL Server, para este caso no se utilizó contraseña
- Nombre del catálogo: se debe elegir la base de datos con el que se almacenó en el servidor, en este caso Edelnor_Prueba2.

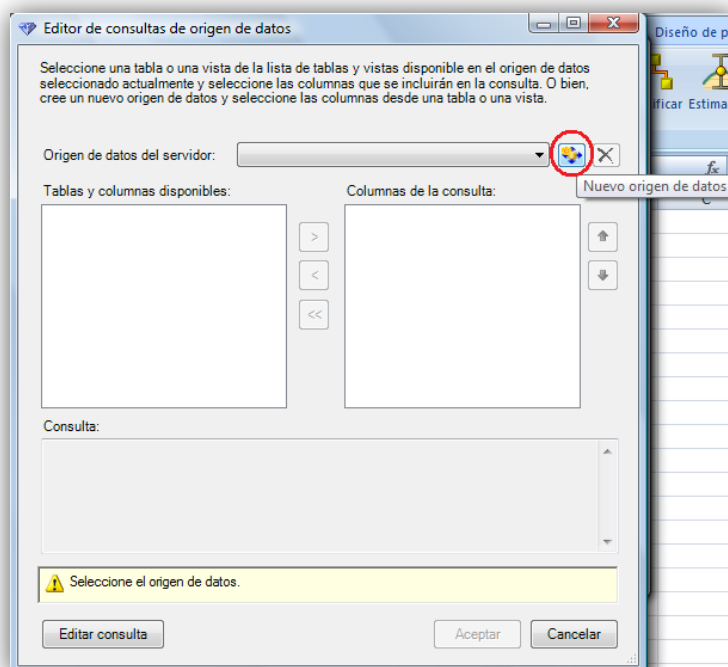


Figura 53: Crear un Nuevo Origen de Datos.

Elaboración: La autora

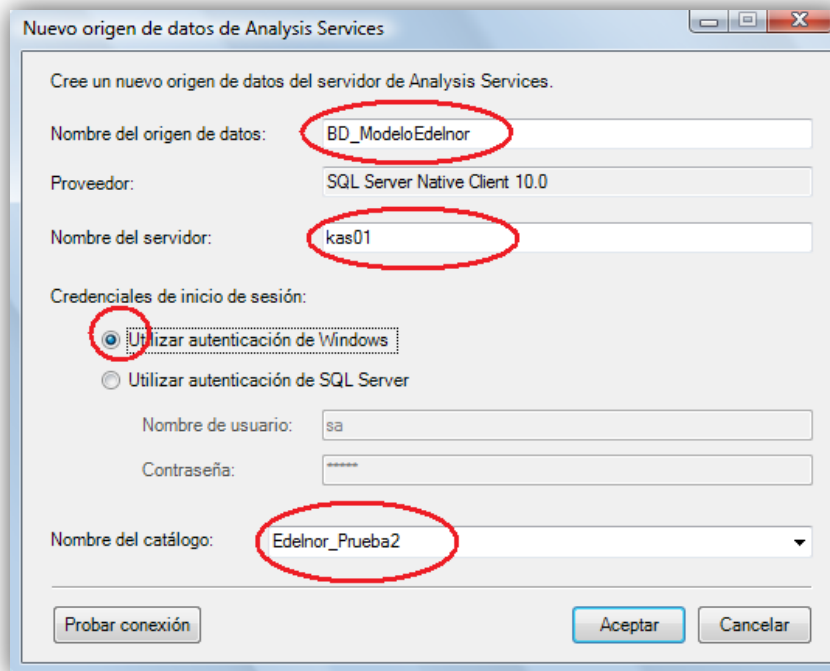


Figura 54: Datos del Servidor de Análisis Services para crear origen de datos.

Elaboración: La autora

Creación de Modelo con el Complemento de Minería de Datos

La creación de modelos tiene por finalidad explicar comportamientos de los sistemas y predecir futuros estados.

Para crear un modelo ingresamos a Microsoft Office Excel 2007 → Minería de datos de datos → Clasificar.

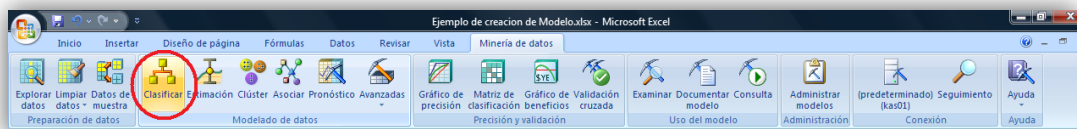


Figura 55: Creación de Modelo – Clasificar.

Elaboración: La autora

La opción Clasificar nos muestra la ventana Asistente para Clasificar, si se desea elegir un intervalo de datos de Excel o datos externos, para este caso elegimos origen de datos externos.

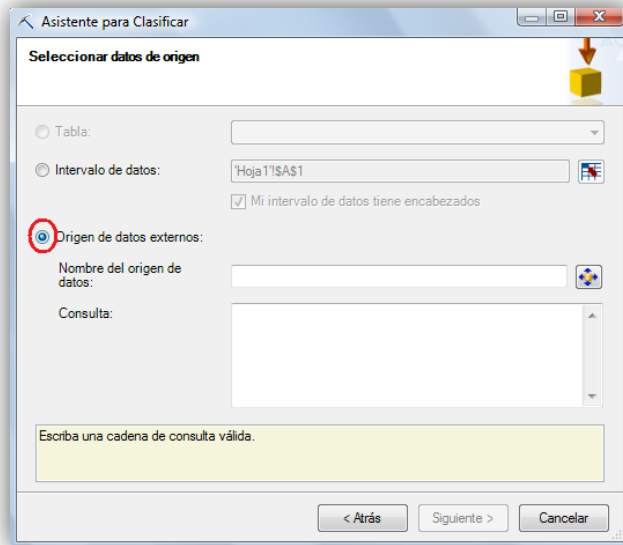


Figura 56: Creación de Modelo - Asistente para Clasificar

Elaboración: La autora

Se elige el origen de datos del servidor que vincula con la base de datos que contiene las vistas y tablas de consulta. Para este caso el origen de datos es BD_ModeloEdelnor y nos muestra las vistas, para este caso, utilizaremos la Vista_Casos_Train_v que contiene los datos de entrada (Número de cuenta de clientes y 12 consumos) .

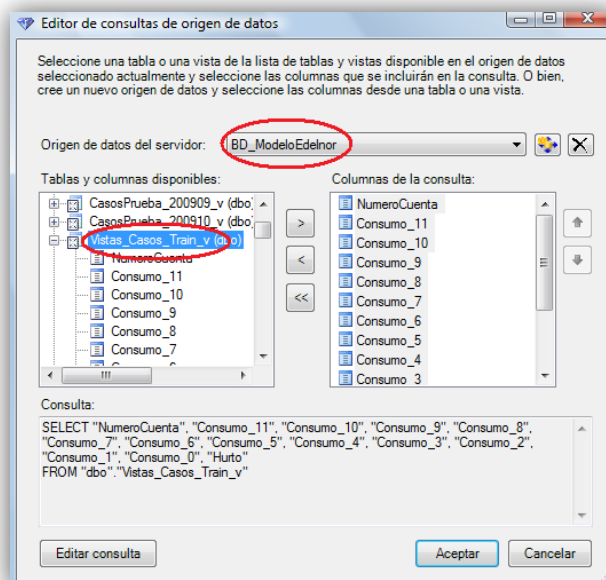


Figura 57: Creación de Modelo - Selección de Datos de Entrenamiento.

Elaboración: La autora

En la ventana Asistente para Clasificar se elige la columna que va a analizar y columnas de entrada, en este caso las columnas de entradas son los 12 consumos y la columna de análisis se elige el Hurto.

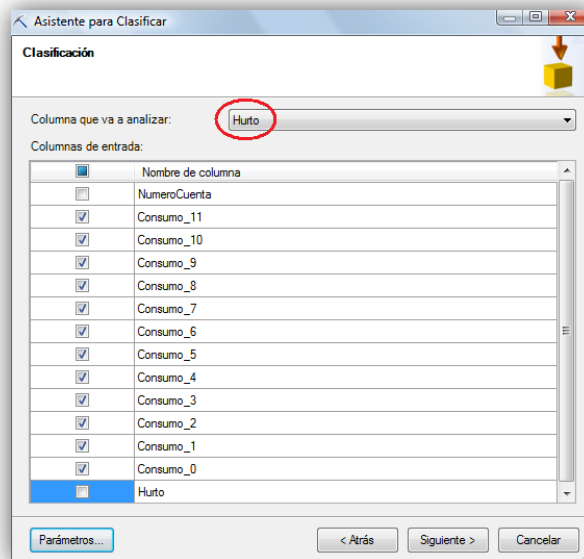


Figura 58: Creación de Modelo - Selección de columna a Analizar.

Elaboración: La autora

Dentro de la ventana de Asistente para Clasificar elegimos Parámetros y nos mostrará la ventana de Parámetros de algoritmo, para este caso elegimos Microsoft Neural Network.

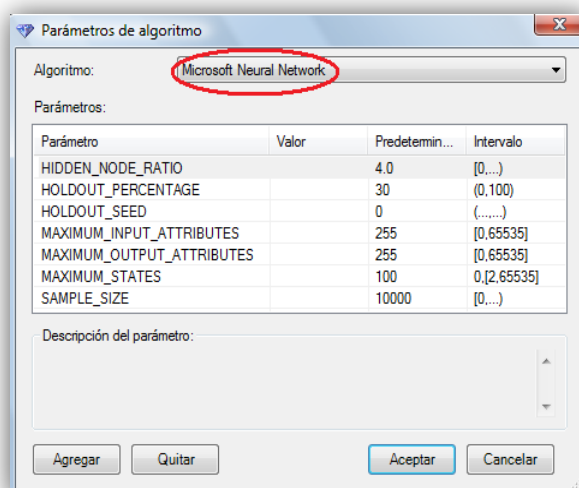


Figura 59: Creación de Modelo - Selección de Algoritmo y parámetros de la Red Neuronal.

Elaboración: La autora

En el asistente para clasificar, elegimos el porcentaje de datos de prueba colocamos 0 porque deseamos que entrene con el 100% de registros seleccionados en la vista.

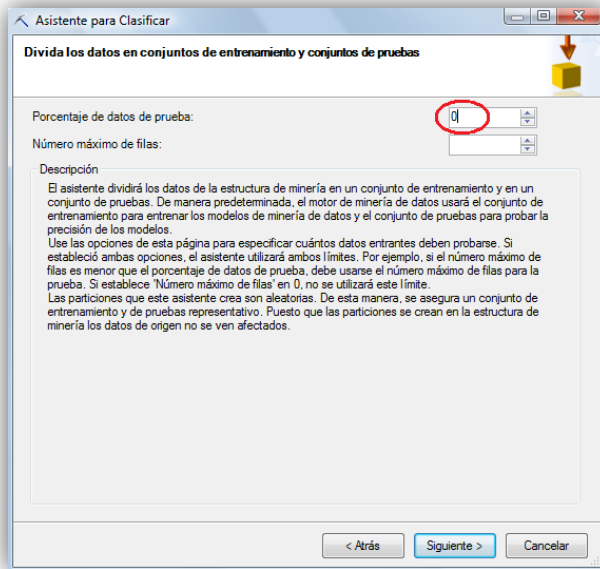


Figura 60: Creación de Modelo - Selección de Datos de Prueba.

Elaboración: La autora

En el asistente para clasificar, colocamos el nombre de la Estructura y nombre del Modelo y finalizamos el proceso de creación del modelo.

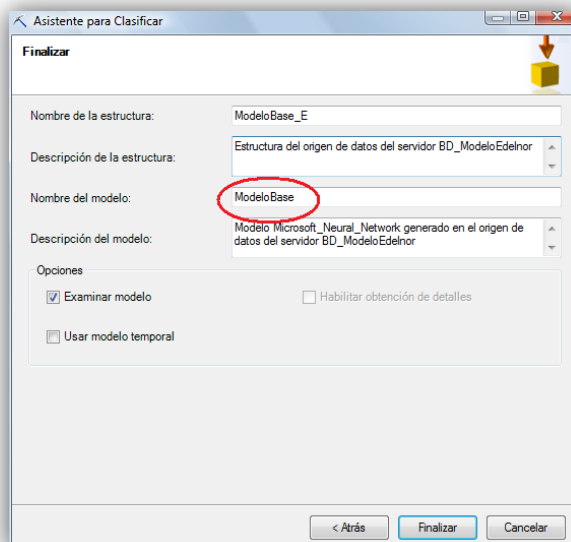


Figura 61: Creación de Modelo - Nombre de Estructura y Modelo.

Elaboración: La autora

Creación de matriz de clasificación del modelo con el complemento de minería de datos

La matriz de clasificación tiene por finalidad permitir evaluar el rendimiento de un modelo existente con respecto a los datos de prueba.

Para crear una matriz de clasificación ingresamos a Microsoft Office Excel → Minería de datos → Matriz de Clasificación.

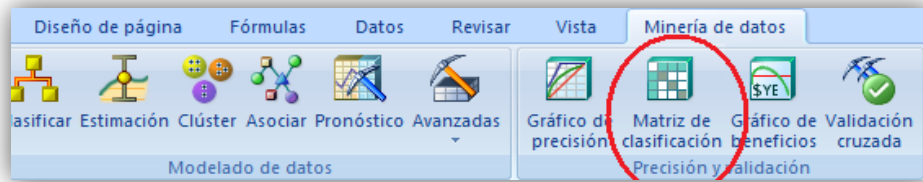


Figura 62: Creación de Matriz de Clasificación.

Elaboración: La autora

Se muestra la ventana de Matriz de clasificación y nos indica seleccionar la estructura y modelo, en este caso elegimos ModeloOptimo1 y al lado derecho como se muestra en figura, se muestra la descripción del modelo (Algoritmo, Salida y Entradas).

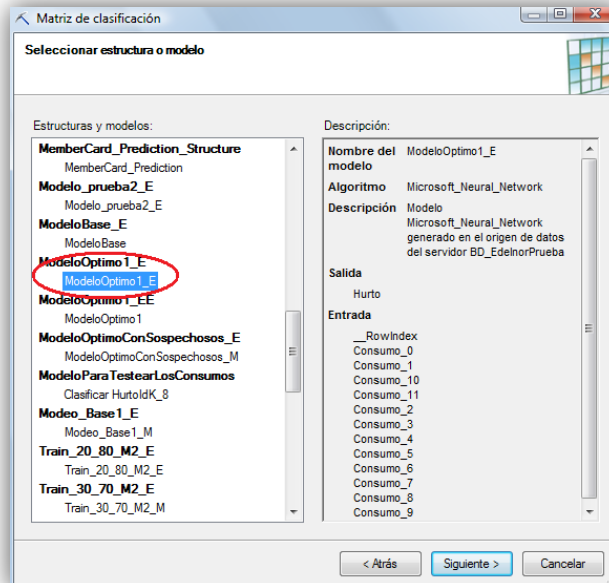


Figura 63: Matriz de Clasificación - Selección de Estructura y Modelo.

Elaboración: La autora

En la ventana Matriz de Clasificación, se especifica la columna que se va a predecir como se muestra en la figura, en este caso, la columna a predecir es el Hurto.

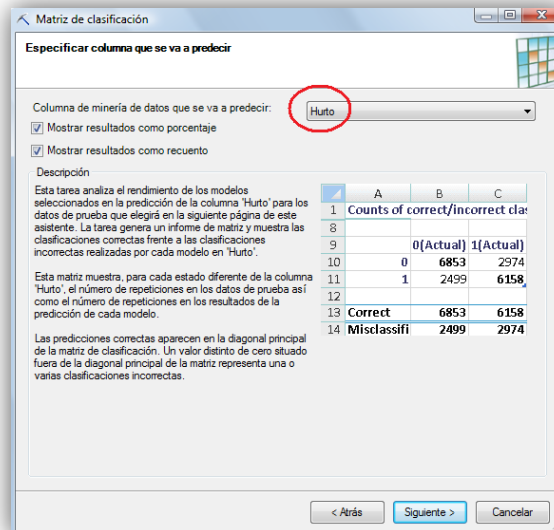


Figura 64: Matriz de Clasificación - Seleccionar columna que se va a predecir.

Elaboración: La autora

Luego se eligen los datos externos a analizar, el origen de datos de donde se extraerá sea reporte, vista o tabla, para este caso se selecciona el origen BD_ModeloBase y la vista Vista_Casos_Test_v.

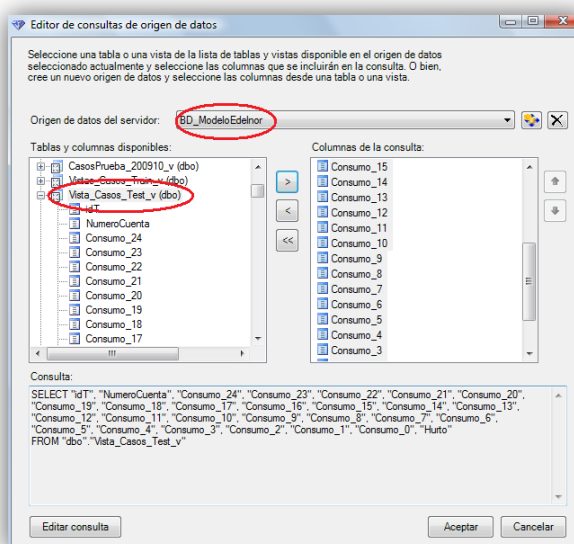


Figura 65: Matriz de Clasificación - Selección de Origen de Datos y Vista.

Elaboración: La autora

Se verifica la relación entre las columnas del modelo y las columnas de la vista a analizar y muestra el resultado en una nueva hoja de excel.

Donde:

Las columnas corresponden a los datos reales.

Las filas corresponden a los datos predichos.

Tabla 35: Descripción de Matriz de Clasificación

	No(Real)	Si(Real)
No	VN (Verdaderos negativos): corresponde a los valores que el modelo predijo correctamente.	FN (Falsos negativos): Corresponde a los valores que el modelo predijo como no hurtadores, pero son hurtadores.
Si	FP (Falsos positivos): Corresponde a los valores que el modelo predijo como hurtadores pero, pero no son hurtadores.	VP (Verdaderos positivos): corresponde a los valores que el modelo predijo correctamente.

Elaboración: La autora

Nombre del modelo:	ModeloOptimo1_M	ModeloOptimo1_M
Total de correctas:	92.28 %	791384
Total de incorrectas:	7.72 %	66179
Resultados como porcentajes para el modelo 'ModeloOptimo1_M'		
	no(Real)	si(Real)
no	95.72 %	55.33 %
si	4.28 %	44.67 %
Correcta	95.72 %	44.67 %
Incorrecta	4.28 %	55.33 %
Resultados como recuentos para el modelo 'ModeloOptimo1_M'		
	no(Real)	si(Real)
no	765591	31952
si	34227	25793
Correcta	765591	25793
Incorrecta	34227	31952

Figura 66: Resultado de la Matriz de Clasificación.

Elaboración: La autora

Evaluación y Consulta del Modelo con el Complemento de Minería de Datos

Consultar un Modelo tiene por finalidad predecir la clase (Hurto o No Hurto) a la que pertenece determinada señal (consumos) en base a los patrones de comportamiento (forma de la curva de consumos de periodos de 12 meses) aprendidos por el modelo (Red Neuronal).

Para consultar un modelo, se debe elegir Microsoft Office Excel 2007 →Minería de Datos→Consulta.

Seleccionar en Minería de Datos la opción Consulta como se muestra en la figura inferior.

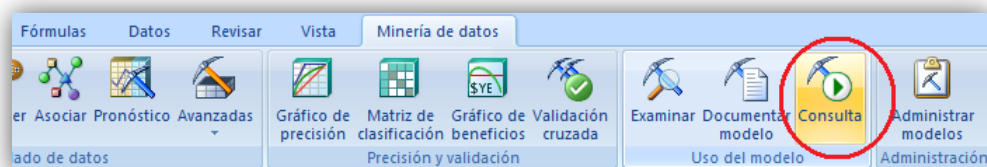


Figura 67: Evaluación de Modelo - Consulta.

Elaboración: La autora

Se muestra la ventana de Asistente para consulta de minería de datos, donde nos indica elegir el Modelo que se desea consultar, en este caso elegimos la estructura ModeloBase_E. como se indica en la figura y a la derecha de la figura vemos los datos del modelo como por ejemplo la salida y las entradas y el tipo de parámetros que se utilizó en su construcción.

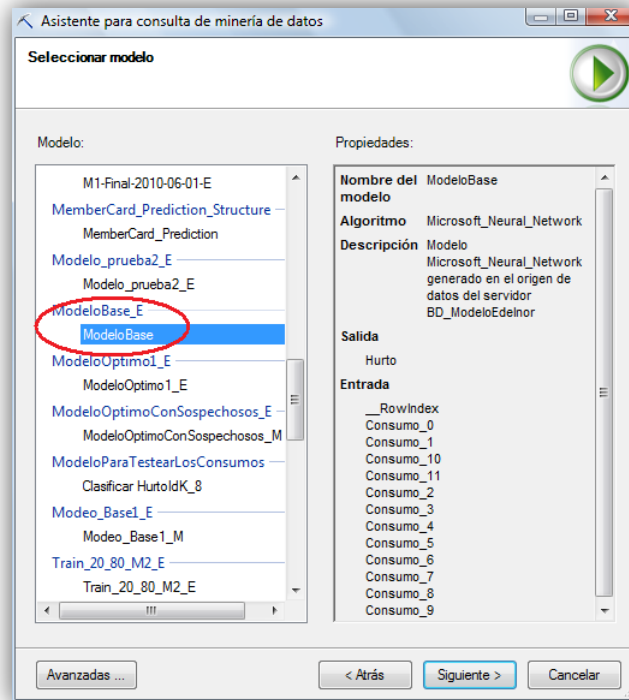


Figura 68: Evaluación de Modelo - Selección de Modelo a Consultar.

Elaboración: La autora

Luego nos muestra el editor de consultas de origen de datos, se elige el Origen de datos del servidor que se creó inicialmente y mostrará las tablas, vistas creadas, para este caso se elige la vista Vista_Casos_Test_v ó datos del reporte y se agrega al lado derecho. Verificar que se agreguen todas las columnas de la consulta.

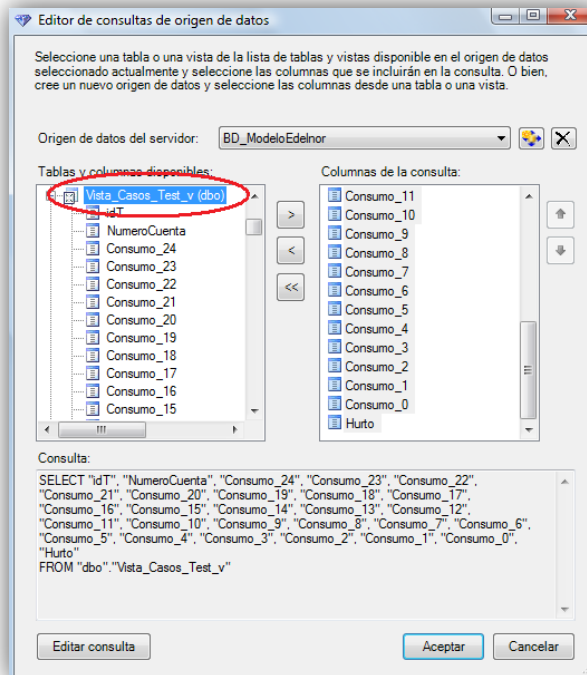


Figura 69: Evaluación de Modelo - Selección de Datos a Evaluar.

Elaboración: La autora

Se verifica la relación de las columnas del modelo seleccionado y los datos externos de consulta (Vista_Casos_Test_v) coincidan como se muestra en la figura.

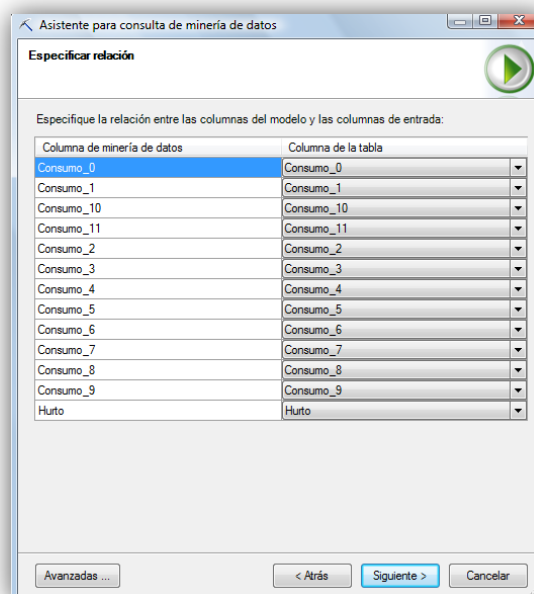


Figura 70: Evaluación de Modelo- Relación de columnas del modelo con la data a consultar. Elaboración: La autora

En el asistente, para consulta de minería de datos agregar salidas, para este caso agregamos Idt, NumeroCuenta, Hurto y Sospecha como se muestran en las figuras.

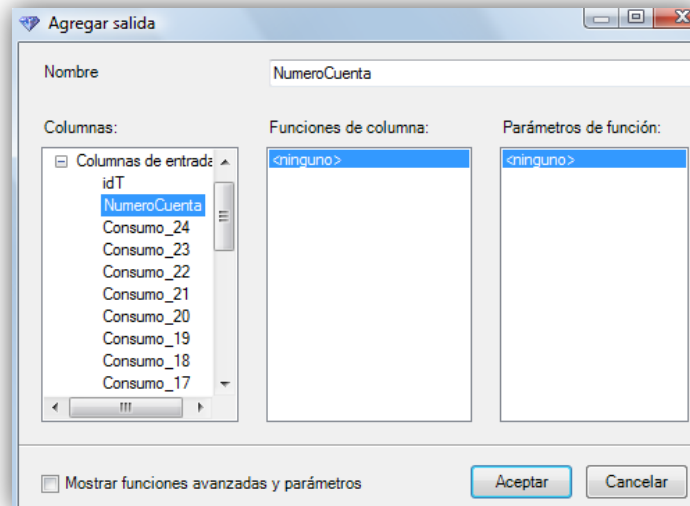


Figura 71: Evaluación de Modelo - Agregar Salida Número de Cuenta del Cliente.

Elaboración: La autora

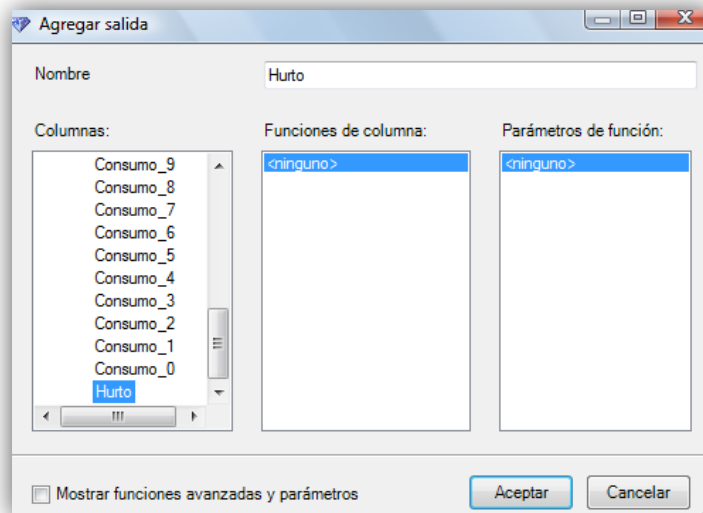


Figura 72: Evaluación de Modelo - Agregar Salida Hurto.

Elaboración: La autora

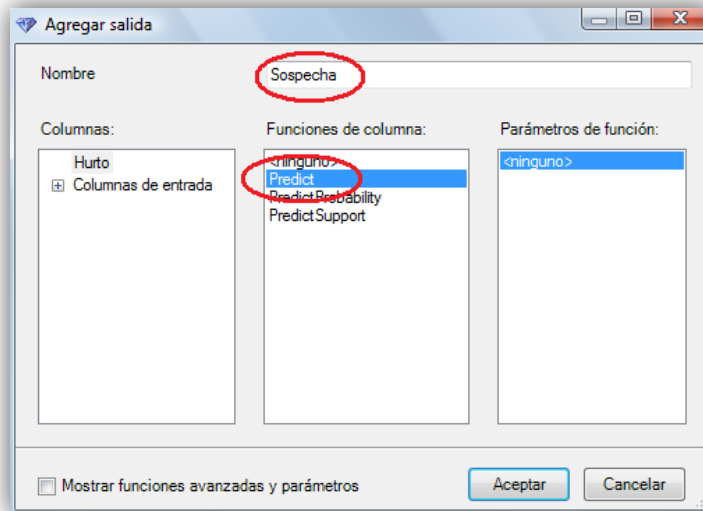


Figura 73: Evaluación del Modelo - Agregar Salida Sospecha.

Elaboración: La autora

En el asistente para consulta de minería de datos, debe elegir una hoja de cálculo donde guardar el resultado, se puede guardar como archivo de Excel o como texto para este caso lo guardamos como Sospecha.xls

idT	NumeroCuenta	Hurto	Sospecha
1	3	no	si
1	88	si	no
2	4	no	si
2	88	si	no
3	5	no	no
3	88	si	no
4	7	no	no
4	88	si	no
5	8	no	no
5	88	si	no
6	9	no	no
6	88	si	no
7	11	no	no
7	88	si	no
8	12	no	no
8	88	si	no
9	13	no	no
9	88	si	no
10	14	no	no
10	88	si	no
11	15	no	no
11	88	si	no
12	16	no	no
12	88	si	no
13	17	no	no

Figura 74: Evaluación de Modelo - Resultado de Evaluación.

Elaboración: La autora

Proceso de optimización

El proceso de optimización tiene como finalidad mejorar la sensibilidad y especificidad de los modelos generados y en este caso, en una diferenciación a priori en los datos de entrenamiento, que permita elegir solo los VN y VP.

Donde:

VP = Población que el modelo detecta correctamente a los fraudulentos (achurada de amarillo).

VN= Población que el modelo detecta correctamente a los no fraudulentos (achurada de rojo).

FN= Población que el modelo detecta incorrectamente a los no fraudulentos (achurada de naranja lado izquierdo).

FP= Población que el modelo detecta incorrectamente a los fraudulentos (achurada de naranja lado derecho).

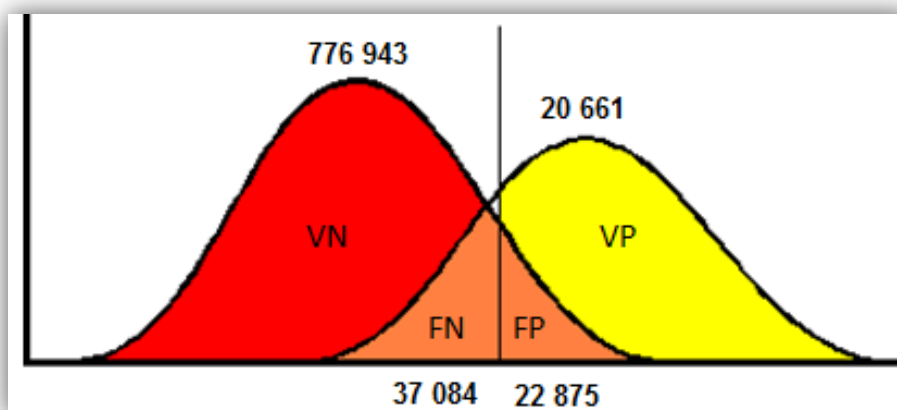


Figura 75: Resultado de Consulta del Modelo

Fuente: Kasperu(2010)

Luego se retira de los datos de entrenamiento la región achurada de naranja (los falsos negativos (FN) y falsos positivos (FP)).

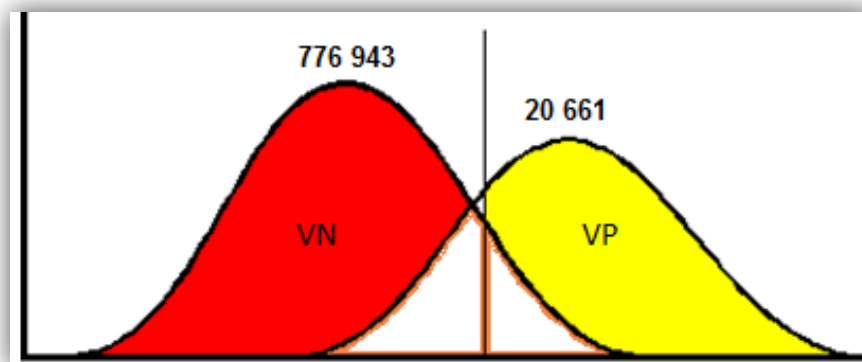


Figura 76: Nuevos datos para Entrenamiento.

Fuente: Kasperu(2010)

Luego para crear el modelo optimizado se debe trabajar solo con los nuevos datos de entrenamiento que contengan solo los VN y VP como se muestra en la figura.

Fase 5: Despliegue

En esta fase ya se ha construido un modelo óptimo (datos más procedimientos) que tiene el mejor desempeño, bajo una perspectiva de análisis de datos. La finalidad de esta etapa es usar el modelo óptimo para predecir comportamientos de nuevos registros de datos que parece tener la alta calidad de una perspectiva de análisis de datos.

Para calcular sospecha de clientes los datos a evaluar sospecha deben tener ciertas características:

- Días de facturación debe ser {28-33}
- Clave de Facturación diferentes de U
- Consumos Mayores a Cero.
- Consumos a evaluar debe ser consumo Promedio Diario.
- Deben Ser Clientes de tipo Habilitado.
- Cliente con Tipo de Tarifa BT5 B (Residenciales)

- Cliente con más de 11 meses de alta.

	NumeroCuenta	Consumo_11	Consumo_10	Consumo_9	Consumo_8	Consumo_7	Consumo_6	Consumo_5	Consumo_4	Consumo_3	Consumo_2	Consumo_1	Consumo_0
1	1867743	8.2000	8.5000	9.1613	10.5161	10.3448	9.4194	10.5806	12.2000	9.6364	9.3214	10.0313	9.9032
2	1654508	14.4063	10.4839	10.7586	10.6129	3.6774	0.0000	2.3548	1.0000	2.5517	1.9355	1.7931	1.5806
3	696826	3.7333	3.7931	4.2000	4.2188	3.7000	3.7241	3.5455	3.5625	3.7419	3.2759	1.5806	0.1935
4	595799	3.7241	3.7576	3.3636	2.7143	2.4643	2.0625	2.3125	2.9485	3.1071	2.7667	2.3939	0.7241
5	1638396	1.0323	4.1212	2.1429	3.8000	3.5758	0.0000	4.0303	3.3571	3.6786	6.0909	9.0938	2.8788
6	1624612	4.9333	5.2727	4.7586	4.6970	5.3929	7.1429	6.6970	6.4375	6.0000	6.8333	8.2903	9.0000
7	1312856	1.8710	2.1379	1.3667	3.5625	4.3333	2.1379	1.7188	1.4000	1.9333	3.4545	1.8214	2.3030
8	1840586	7.6970	5.7857	8.2500	7.6563	7.0909	4.6667	2.3000	0.2667	0.0000	0.0000	0.0000	2.7333
9	1005182	8.6333	10.0000	9.5000	6.0938	5.1000	5.2759	5.2121	5.0938	5.4516	3.0345	5.0323	6.1935
10	454838	2.6333	2.5667	2.6667	2.4839	2.5161	2.4483	2.6774	2.7419	2.7000	2.6970	2.6786	2.7188

Figura 77: Datos a evaluar sospechoso.

Elaboración: La autora

Pasos para que el usuario consulte el modelo

Paso 1: Debe ingresar a la opción “Consulta”



Figura 78: Complemento Minería de Datos – Opción Consulta

Elaboración: La autora

Paso 2: El usuario debe seleccionar el modelo a consultar.

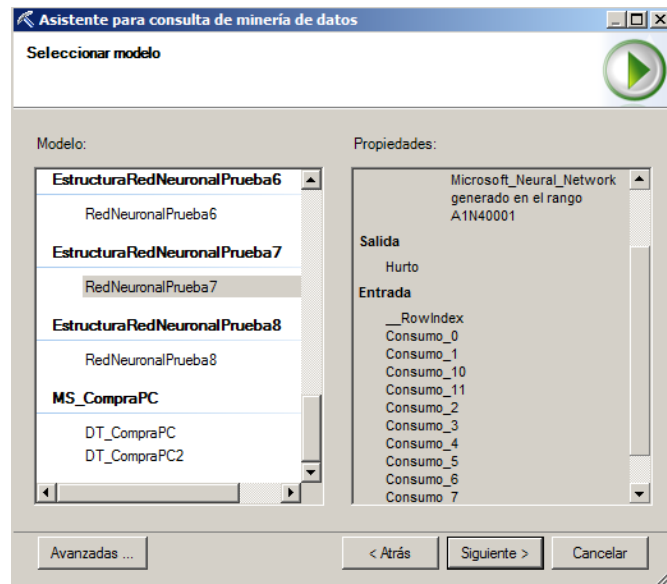


Figura 79: Complemento Minería de Datos – Opción Asistente de Consulta

Elaboración: La autora

Paso 3: El usuario debe seleccionar los datos a consultar de acuerdo con la estructura del modelo Clave y 12 consumos a consultar sospechosos.

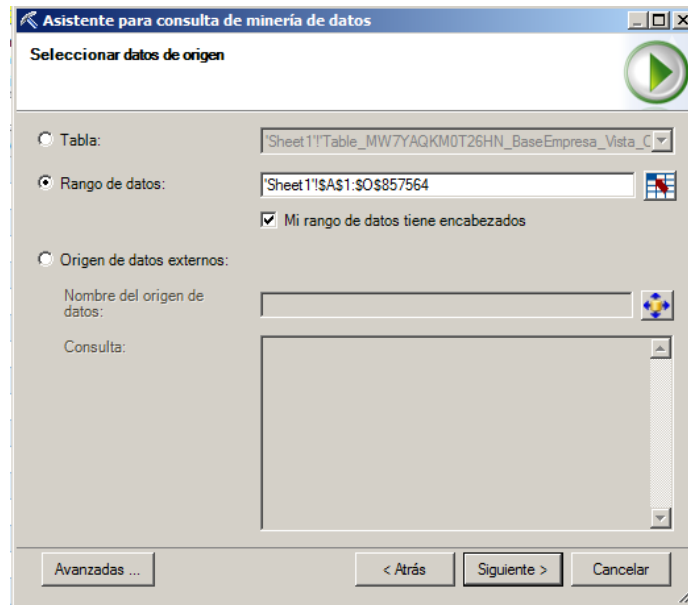


Figura 80: Complemento Minería de Datos – Opción Datos Origen

Elaboración: La autora

Paso 4: El usuario debe comparar que los datos a evaluar sospecha coincide con la estructura del modelo a consultar.

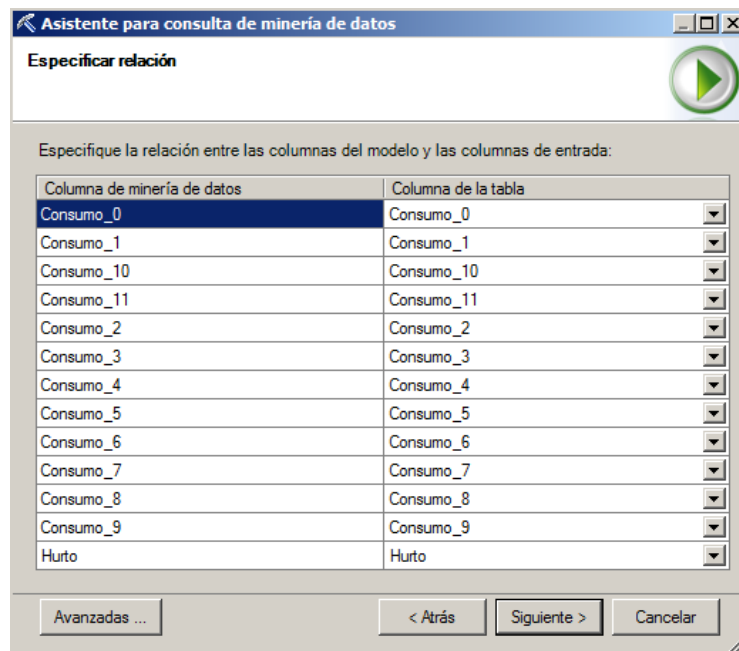


Figura 81: Complemento Minería de Datos – Opción relación de columnas

Elaboración: La autora

Paso 5: El usuario agrega los campos a consultar de cada consumidor la probabilidad de hurto.

idT	NumeroCuenta	Hurto Real	Prediccion	Prob SI	Prob No
1	3	no	si	0.71115821	0.28874182
1	88	si	no	0.329875754	0.670024276
2	4	no	si	0.810675644	0.189224386
2	88	si	no	0.366376107	0.633523923
3	5	no	no	0.396891275	0.603008755
3	88	si	no	0.370952324	0.628947706
4	7	no	si	0.511634465	0.488265565
4	88	si	no	0.343390687	0.656509343
5	8	no	no	0.469580423	0.530319607
5	88	si	no	0.382659681	0.617240349
6	9	no	no	0.487232888	0.512667142
6	88	si	no	0.377923507	0.621976523
7	11	no	si	0.558491351	0.441408679
7	88	si	no	0.347969266	0.651930764
8	12	no	no	0.438985477	0.560914553
8	88	si	no	0.386494458	0.613405572
9	13	no	no	0.495461905	0.504438125

Figura 82: Complemento Minería de Datos – Resultado de Consulta

Elaboración: La autora

Cantidad de Energía detectada por el modelo calculada en S/.

De la consulta, elegimos los clientes que el modelo predijo correctamente, si evaluamos solo energía detectada en el mes:

Tabla 36: Resultado de Cantidad de Energía Detectada.

Sospecha Correcta	Periodo	Cantidad de Energía detectada KW	Costo x KW Aprox. Clientes residenciales Sin igv	Costo total Detectado S/.
10399	Corresponde Un Mes	3015739.1143	0.38	1'145,980.863434
10399	Corresponde a n Meses atrás	9091113.6406	0.38	3'454,623.183428

Elaboración: La autora



Sospechoso.xlsx

CAPÍTULO IV

PRUEBAS Y RESULTADOS

4.1 Las pruebas tienen por finalidad probar los modelos creados y evaluar los resultados para elegir el de mejor desempeño.

Pruebas a realizar

Creación de modelo con datos desbalanceados

Descripción de Pruebas 01:

Para la creación del modelo, se extrajeron intervalos de 12 meses, el periodo a predecir es 200910.

Datos entrenamiento: 10000 casos de hurtos con aplicación de técnica de ventaneo y mayores con promedio mayor a 11 KW y 20000 casos de hurto que nunca hurtaron y que nunca hayan tenido alguna irregularidad con promedio de consumo de los 12 meses mayor a a 11 KW.

Datos de evaluación: 12 consumos de 857563 casos de clientes del periodo 200910.

Ejemplo de vista de consulta:

```
create view dbo.Vistas_Casos_Train_Prueba1_v as

SELECT      COUNT(*)
FROM        Consulta
WHERE       HurtoIdk = 'si' and Consumo_11+ Consumo_10+ Consumo_9+
Consumo_8+ Consumo_7+ Consumo_6+ Consumo_5+ Consumo_4+ Consumo_3+
Consumo_2+
           Consumo_1+ Consumo_0 >11
           ORDER BY NewId()

UNION
SELECT      TOP 20000 C.NumeroCuenta, C.Consumo_11, C.Consumo_10,
C.Consumo_9, C.Consumo_8, C.Consumo_7, C.Consumo_6, C.Consumo_5,
C.Consumo_4,
           C.Consumo_3, C.Consumo_2, C.Consumo_1,
C.Consumo_0, HurtoIdk AS Hurto
FROM        Consulta C, BaseConsumo B
WHERE       HurtoIdk = 'no' AND C.NumeroCuenta = B.NumeroCuenta AND
B.CodigoIrregularidad_Ult_Deteccion IS NULL and c.Consumo_11+
c.Consumo_10+ c.Consumo_9+ c.Consumo_8+ c.Consumo_7+ c.Consumo_6+
c.Consumo_5+ c.Consumo_4+ c.Consumo_3+ c.Consumo_2+
           c.Consumo_1+ c.Consumo_0 >11
           ORDER BY NewId()
```

Resultado de la evaluación

El resultado de la prueba realizada no fue exitosa debido a que el modelo detecta excelente los casos de No Hurto con un 97.34% y los casos de Hurto con un 23.08%.

Ejemplo de vista de evaluación:

```
SELECT idT, NumeroCuenta, Consumo_24, Consumo_23, Consumo_22, Consumo_21,
Consumo_20, Consumo_19, Consumo_18, Consumo_17, Consumo_16, Consumo_15,
Consumo_14, Consumo_13, Consumo_12, Consumo_11, Consumo_10, Consumo_9,
Consumo_8, Consumo_7, Consumo_6, Consumo_5, Consumo_4, Consumo_3, Consumo_2,
Consumo_1, Consumo_0, HurtoIdK AS Hurto
```

```
FROM      dbo.Consulta
```

Columna de predicción 'Hurto'			
Las columnas corresponden a valores reales			
Las filas corresponden a valores predichos			
Nombre del modelo:	ModeloRedNeuronalPrueba1	ModeloRedNeuronalPrueba1	
Total de correctas:	97.30 %	97302	
Total de incorrectas:	2.70 %	2698	
Resultados como porcentajes para el modelo 'ModeloRedNeuronalPrueba2'			
	no(Real)	si(Real)	
no	97.34 %	76.92 %	
si	2.66 %	23.08 %	
Correcta	97.34 %	23.08 %	
Incorrecta	2.66 %	76.92 %	
Resultados como recuentos para el modelo 'ModeloRedNeuronalPrueba2'			
	no(Real)	si(Real)	
no	97290	40	
si	2658	12	
Correcta	97290	12	
Incorrecta	2658	40	

Figura 83: Resultado de Prueba 01

Elaboración: La autora

Descripción de pruebas 02:

Para la creación del modelo, se extrajeron intervalos de 12 meses, el periodo a predecir es 200910.

Datos entrenamiento: 10000 casos de hurtos con aplicación de técnica de ventaneo y mayores con promedio mayor a 11 KW y 20000 casos de hurto que nunca hurtaron y que nunca hayan tenido alguna irregularidad.

Datos de evaluación: 12 consumos de 100000 casos de clientes aleatorios del periodo 200910.

Ejemplo de vista de consulta:

```

create view dbo.Vistas_Casos_Train_Prueba2_v as

SELECT      COUNT(*)
FROM        Consulta
WHERE       HurtoIdk = 'si' and Consumo_11+ Consumo_10+ Consumo_9+
Consumo_8+ Consumo_7+ Consumo_6+ Consumo_5+ Consumo_4+ Consumo_3+
Consumo_2+
           Consumo_1+ Consumo_0 >11
           ORDER BY NewId()

UNION
SELECT      TOP 20000 C.NumeroCuenta, C.Consumo_11, C.Consumo_10,
C.Consumo_9, C.Consumo_8, C.Consumo_7, C.Consumo_6, C.Consumo_5,
C.Consumo_4,
           C.Consumo_3, C.Consumo_2, C.Consumo_1,
C.Consumo_0, HurtoIdk AS Hurto
FROM        Consulta C, BaseConsumo B

```

```

WHERE HurtoIdk = 'no' AND C.NumeroCuenta = B.NumeroCuenta AND
B.CodigoIrregularidad_Ult_Deteccion IS NULL
ORDER BY NewId()

```

Resultado de la evaluación

El resultado de la prueba realizada no fue exitosa debido a que el modelo detecta excelente los casos de No Hurto con un 97.27% y los casos de Hurto con un 21.15%, también se visualiza que por extraer aleatorios trajo poco casos de hurto para evaluar.

Ejemplo de vista de evaluación:

```

SELECT top 100000 NumeroCuenta, Consumo_24, Consumo_23, Consumo_22,
Consumo_21, Consumo_20, Consumo_19, Consumo_18, Consumo_17, Consumo_16,
Consumo_15, Consumo_14, Consumo_13, Consumo_12, Consumo_11,
Consumo_10, Consumo_9, Consumo_8, Consumo_7, Consumo_6, Consumo_5,
Consumo_4, Consumo_3, Consumo_2, Consumo_1, Consumo_0, HurtoId AS
Hurto FROM dbo.Maestro order by newid()

```

1 Recuentos de clasificación correcta o incorrecta para el modelo 'Modelo			
2 Columna de predicción 'Hurto'			
3 Las columnas corresponden a valores reales			
4 Las filas corresponden a valores predichos			
5			
6	Nombre del modelo:	ModeloRedNeuronalPrueba3	ModeloRedNeuronalPrueba3
7	Total de correctas:	97.23 %	97234
8	Total de incorrectas:	2.77 %	2766
9			
10 Resultados como porcentajes para el modelo 'ModeloRedNeuronalPrueba3'			
11		no(Real)	si(Real)
12	no	97.27 %	78.85 %
13	si	2.73 %	21.15 %
14			
15	Correcta	97.27 %	21.15 %
16	Incorrecta	2.73 %	78.85 %
17			
18 Resultados como recuentos para el modelo 'ModeloRedNeuronalPrueba3'			
19		no(Real)	si(Real)
20	no	97223	41
21	si	2725	11
22			

Figura 84: Resultado de Prueba 02

Elaboración: La autora

Creación de Modelo con datos balanceados

Descripción de Pruebas 01:

Para la creación del modelo, se extrajeron intervalos de 12 meses, el periodo a predecir es 200910.

Datos entrenamiento: 20000 casos de hurtos con aplicación de técnica de ventaneo y mayores con promedio mayor a 11 KW y 20000 casos de hurto y que nunca hayan tenido alguna irregularidad con promedio de consumo de los 12 meses mayor a a 11 KW.

Ejemplo vista de Entrenamiento:

```
create view dbo.Vistas_Casos_Train_Prueba7_v as
```

```
SELECT TOP 20000 NumeroCuenta, Consumo_11, Consumo_10, Consumo_9,  
Consumo_8, Consumo_7, Consumo_6, Consumo_5, Consumo_4, Consumo_3, Consumo_2,  
Consumo_1, Consumo_0, HurtoIdk AS Hurto FROM Consulta
```

```
WHERE HurtoIdk = 'si' and Consumo_11+ Consumo_10+ Consumo_9+ Consumo_8+  
Consumo_7+ Consumo_6+ Consumo_5+ Consumo_4+ Consumo_3+ Consumo_2+  
Consumo_1+ Consumo_0 >11
```

```
ORDER BY NewId()
```

```
UNION
```

```
SELECT TOP 20000 C.NumeroCuenta, C.Consumo_11, C.Consumo_10, C.Consumo_9,  
C.Consumo_8, C.Consumo_7, C.Consumo_6, C.Consumo_5, C.Consumo_4, C.Consumo_3,  
C.Consumo_2, C.Consumo_1, C.Consumo_0, HurtoIdk AS Hurto FROM Consulta C,  
BaseConsumo B
```

```
WHERE HurtoIdk = 'no' AND C.NumeroCuenta = B.NumeroCuenta AND  
B.CodigoIrregularidad_Ult_Deteccion IS NULL and c.Consumo_11+ c.Consumo_10+  
c.Consumo_9+ c.Consumo_8+ c.Consumo_7+ c.Consumo_6+ c.Consumo_5+  
c.Consumo_4+ c.Consumo_3+ c.Consumo_2+ c.Consumo_1+ c.Consumo_0 >11 ORDER  
BY NewId()
```

Datos de evaluación: 12 consumos de 857563 casos de clientes del periodo 200910.

Ejemplo de vista de consulta:

```
SELECT idT, NumeroCuenta, Consumo_24, Consumo_23, Consumo_22, Consumo_21,  
Consumo_20, Consumo_19, Consumo_18, Consumo_17, Consumo_16, Consumo_15,  
Consumo_14, Consumo_13, Consumo_12, Consumo_11, Consumo_10, Consumo_9,  
Consumo_8, Consumo_7, Consumo_6, Consumo_5, Consumo_4, Consumo_3, Consumo_2,  
Consumo_1, Consumo_0, HurtoIdK AS Hurto
```

FROM dbo.Consulta

Resultado de la evaluación

El resultado de la prueba realizada fue exitosa debido a que el modelo detecta muy bien los casos de No Hurto con un 89.29% (De 799 818 detecta correctamente a 714 595 que no son hurtadores) y los casos de Hurto con un 55.03% (De 57745 detecta correctamente a 31779 hurtadores) y los falsos positivos es bajo 10.71 % corresponden a casos que el modelo dice que son hurtadores pero realmente no lo son, hasta el momento es el modelo más acertado de las pruebas realizadas.

4	Las filas corresponden a valores predichos		
5			
6	Nombre del modelo:	RedNeuronalPrueba7	RedNeuronalPrueba7
7	Total de correctas:	86.99 %	745974
8	Total de incorrectas:	13.01 %	111589
9			
10	Resultados como porcentajes para el modelo 'RedNeuronalPrueba7'		
11		no(Real)	si(Real)
12	no	89.29 %	44.97 %
13	si	10.71 %	55.03 %
14			
15	Correcta	89.29 %	55.03 %
16	Incorrecta	10.71 %	44.97 %
17			
18	Resultados como recuentos para el modelo 'RedNeuronalPrueba7'		
19		no(Real)	si(Real)
20	no	714195	25966
21	si	85623	31779
22			
23	Correcta	714195	31779
24	Incorrecta	85623	25966

Figura 85: Resultado de Prueba 03

Elaboración: La autora

Descripción de pruebas 02:

Para la creación del modelo se utilizó se extrajo intervalos de 12 meses, el periodo a predecir es 200910.

Datos entrenamiento: 30000 casos de hurtos con aplicación de técnica de ventaneo y mayores con promedio mayor a 11 KW y 20000 casos de hurto que nunca hurtaron y que nunca hayan tenido alguna irregularidad con promedio de consumo de los 12 meses mayor a a 11 KW.

Ejemplo vista de entrenamiento:

```
create view dbo.Vistas_Casos_Train_Prueba8_v as
```

```
SELECT TOP 30000 NumeroCuenta, Consumo_11, Consumo_10, Consumo_9,  
Consumo_8, Consumo_7, Consumo_6, Consumo_5, Consumo_4, Consumo_3, Consumo_2,  
Consumo_1, Consumo_0, HurtoIdk AS Hurto FROM Consulta
```

```
WHERE HurtoIdk = 'si' and Consumo_11+ Consumo_10+ Consumo_9+ Consumo_8+  
Consumo_7+ Consumo_6+ Consumo_5+ Consumo_4+ Consumo_3+ Consumo_2+  
Consumo_1+ Consumo_0 >11
```

```
ORDER BY NewId()
```

```
UNION
```

```
SELECT TOP 20000 C.NumeroCuenta, C.Consumo_11, C.Consumo_10, C.Consumo_9,  
C.Consumo_8, C.Consumo_7, C.Consumo_6, C.Consumo_5, C.Consumo_4, C.Consumo_3,  
C.Consumo_2, C.Consumo_1, C.Consumo_0, HurtoIdk AS Hurto FROM Consulta C,  
BaseConsumo B
```

```
WHERE HurtoIdk = 'no' AND C.NumeroCuenta = B.NumeroCuenta AND  
B.CodigoIrregularidad_Ult_Deteccion IS NULL and c.Consumo_11+ c.Consumo_10+  
c.Consumo_9+ c.Consumo_8+ c.Consumo_7+ c.Consumo_6+ c.Consumo_5+  
c.Consumo_4+ c.Consumo_3+ c.Consumo_2+ c.Consumo_1+ c.Consumo_0 >11 ORDER  
BY NewId()
```

Datos de evaluación: 12 consumos de 857563 casos de clientes del periodo 200910.

Ejemplo de vista de consulta:

```
SELECT idT, NumeroCuenta, Consumo_24, Consumo_23, Consumo_22, Consumo_21,  
Consumo_20, Consumo_19, Consumo_18, Consumo_17, Consumo_16, Consumo_15,  
Consumo_14, Consumo_13, Consumo_12, Consumo_11, Consumo_10, Consumo_9,  
Consumo_8, Consumo_7, Consumo_6, Consumo_5, Consumo_4, Consumo_3, Consumo_2,  
Consumo_1, Consumo_0, HurtoIdK AS Hurto
```

```
FROM dbo.Consulta
```

Resultado de la evaluación

Al aumentar el número de casos de hurto el resultado de la prueba realizada no fue exitosa debido a que el modelo detecta muy bien a los casos de Hurto con un 85.05% mientras que los casos de no hurto son 38.19%.

Recuentos de clasificación correcta o incorrecta para el modelo			
Columna de predicción 'Hurto'			
Las columnas corresponden a valores reales			
Las filas corresponden a valores predichos			
Nombre del modelo:	RedNeuronalPrueba8	RedNeuronalPrueba8	
Total de correctas:	41.35 %		354581
Total de incorrectas:	58.65 %		502982
Resultados como porcentajes para el modelo 'RedNeuronalPrueba8'			
	no(Real)	si(Real)	
no	38.19 %		14.95 %
si	61.81 %		85.05 %
Correcta	38.19 %		85.05 %
Incorrecta	61.81 %		14.95 %
Resultados como recuentos para el modelo 'RedNeuronalPrueba8'			
	no(Real)	si(Real)	
no	305467		8631
si	494351		49114
Correcta	305467		49114
Incorrecta	494351		8631

Figura 86: Resultado de Prueba 04

Elaboración: La autora

Cuadro Resumen de Resultados de Pruebas realizadas alineados a objetivos.

El presente cuadro corresponde a las pruebas realizadas en la creación del modelo y los datos utilizados en su evaluación.

La sensibilidad indica la probabilidad de que el modelo detecte correctamente a los casos de fraude.

Las especificidad indica la probabilidad de que el modelo detecte correctamente a los casos de no fraude.

Tabla 37: Cuadro de Pruebas y Resultados

NOMBRE MODELO	DESCRIPCIÓN	DATOS DE PRUEBA	RESULTADO	SENSIBILIDAD	ESPECIFICIDAD	VP	FP	VN	FN	Éxito	Error
ModeloRN_01	Entrenamiento con datos de 12 meses con datos desbalanceados del periodo 200910 con 10000 casos de hurtos con consumos mayores con promedio mayor a 11 KW y 20000 casos de hurto que nunca hurtaron con promedio mayor a 11 KW.	El 100 000 casos aleatorios del periodo 200910	El resultado de la prueba realizada no fue exitosa debido a que el modelo detecta excelente los casos de No Hurto con un 97.34% y los casos de Hurto con un 23.08%.	23.08%	97.34%	12	2658	97290	40	97.30%	2.70%
ModeloRN_02	Entrenamiento con datos de 12 meses con datos desbalanceados 20000 casos de hurtos con promedio mayor a 11 KW y 30000 casos de No hurto promedio de consumo mayor a 11 KW.	Con 857563 casos de clientes del periodo 200910.	El resultado de la prueba fue parcialmente exitosa debido a que el modelo detecta muy bien los casos de No Hurto con un 92.72% y los casos de Hurto con un 48.89%.	48.89%	92.72%	31395	58468	744242	32822	89.47%	10.53%
ModeloRN_03	Entrenamiento con datos de 12 meses con datos balanceados 20000 casos de hurtos y con promedio mayor a 11 KW y 20000 casos de No hurto con promedio de consumo mayor a 11 KW.	Con 857563 casos de clientes del periodo 200910.	El resultado de la prueba realizada fue exitosa debido a que el modelo detecta muy bien los casos de No Hurto con un 89.29% y los casos de Hurto con un 55.03%.	55.03%	89.29%	31779	85623	714195	25966	86.99%	13.01%
ModeloRN_04	Entrenamiento con datos de 12 meses con datos balanceados normalizados con 20000 casos de hurtos y 20000 casos de No hurto.	Con 857563 casos de clientes del periodo 200910.	El resultado de la prueba no fue exitosa debido a que el modelo detecta muy bien los casos de Hurto con un 63.64% y los casos de No Hurto con un 26.33%.	63.64%	26.33%	22040	554183	198077	12591	27.97%	72.03%
ModeloRN_05	Entrenamiento con datos de 12 meses con datos desbalanceados normalizados con 20000 casos de hurtos y 30000 casos de No hurto	Con 857563 casos de clientes del periodo 200910.	El resultado de la prueba no fue exitosa debido a que el modelo detecta muy bien los casos de No Hurto con un 67.88% y los casos de Hurto con un 40.42%.	40.42%	67.88%	13999	241610	510650	20632	66.67%	33.33%

Elaboración: La autora

CONCLUSIONES

- 1 El modelo desarrollado permite evaluar y detectar las fraudes de energía eléctrica con un alto grado de sensibilidad 55.03% y especificidad es de 89% en clientes residenciales, siendo su valor de éxito el 86.99% mientras que el proceso tradicional según la memoria anual publicada por la empresa eléctrica Edelnor en el 2011 se realizaron 193,516 inspecciones se logra regularizar el 17419 consumos no registrados equivalente al 9% porcentaje de éxito.
- 2 Las redes neuronales se utilizan para el aprendizaje de comportamiento basado en la historia de casos de fraude y no fraude, queda demostrado su comportamiento predictivo con el resultado de pruebas donde se logró detectar satisfactoriamente los casos de fraude con datos balanceados 20000(50%) casos de hurto y 20000(50%) casos de no hurto.
- 3 El aumentar los casos de hurto en el entrenamiento del modelo no garantiza que el modelo sea óptimo, por ello se recomienda trabajar con datos balanceados.
- 4 La normalización de datos para aprender los patrones de comportamientos mejora la sensibilidad del modelo pero no es bueno aprendiendo los casos de no hurto.
- 5 La preparación de los datos para crear los modelos y obtener los patrones del comportamiento tomó el 60% de esfuerzo de todo el proyecto y se tuvo que regresar a una etapa anterior para mejorar las pruebas.

RECOMENDACIONES

- 1 El modelo de detección de fraude se realizó sobre energía eléctrica pero también se sugiere realizar el mismo proceso para detectar pérdidas en otros servicios agua, gas, etc.
- 2 Para consultar sospechosos de hurto, en la actualidad, se recomienda volver a generar un modelo con los pasos y procedimientos establecidos en el proyecto pero con información actualizada de un periodo anterior hacia atrás 25 meses del periodo que se desea predecir.
- 3 Si se desea desarrollar un nuevo modelo minería de datos, es recomendable apoyarse en expertos del negocio para determinar el conjunto de entrenamiento y prueba para un mejor resultado.

FUENTES DE INFORMACIÓN

Bibliográficas

1. Berry M. y Linoff G. (2004) Data Mining Techniques: for marketing, sales and customer relationship management.
2. Ander-Egg Ezequiel (1995). Técnicas de investigación Social. Editorial LUMEN. 24ª Edición. Buenos Aires. Argentina obtenida el 10 de mayo del 2014 desde http://www.franciscohuertas.com.ar/wp-content/uploads/2011/04/IT_Ander-Egg_1.pdf
3. García Julio y Martínez (2000). El debate investigación cualitativa frente a investigación cuantitativa. Universidad de Alicante. Dpto. Enfermería Clínica, Vol. 6 Núm. 5
4. Knowledge And System Perú SAC - KASPERU (2010) Especialización en Análisis Predictivo y Minería de datos, documentos no publicado.
5. Larose, T.(2005) Data Mining Methods and Models. New York. Editorial John Wiley & Sons, Inc.
6. Moreno Benjamín (2009). Minería sobre cantidades de datos. Tesis de grado para obtener el grado de Maestro en Ciencia de la facultad de

Ciencias y Tecnología de la Información) obtenido el 26 de Junio de 2014.

7. Rodríguez Miguel y Llorca Javier (2004). Estudios Longitudinales: Concepto y Particularidades. Universidad de Jaèn. Universidad de Cantabria.

Electrónicas

1. Acevedo Parra, Jorge Luis & Sánchez Calderón, Edgar Alejandro. Perdidas No técnicas: Uso del equipo detector de derivaciones para la identificación de tomas ilegales en acometidas. *Afinidad Eléctrica* [Revista en Línea] IEEE PES Transmission and Distribution Conference and Exposition Latin America. Venezuela 2006 Disponible en: <http://www.afinidadelectrica.com.ar/articuloscat.php?cat=perdidas>
2. Basogain Xavier (2008). *Redes Neuronales artificiales y sus aplicaciones*. Obtenido 01 de mayo del 2014 disponible en: http://cvb.ehu.es/open_course_ware/castellano/tecnicas/redes_neuro/contenidos/pdf/libro-del-curso.pdf
3. Cañar Olmedo, Santiago Patricio (2007). Cálculo Detallado de Pérdidas en Sistemas Eléctricos de Distribución Aplicado al Alimentador "Universidad" Perteneciente a la Empresa Eléctrica Ambato Regional Centro Norte S.A.[Tesis Previa a la Obtención del Título de Ingeniero Eléctrico]. Facultad de Ingeniería Eléctrica y Electrónica. Escuela Politécnica Nacional. Quito. Pp. 22. Obtenido el 10 de abril del 2014 disponible en: <http://bibdigital.epn.edu.ec/bitstream/15000/4217/1/CD-0926.pdf>
4. Cravero Ania, Sepúlveda Samuel (2009). Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio. Universidad de la Frontera. Dpto. Ingeniería de Sistemas, Temuco, Chile. Página 1

<file:///C:/Users/j.flores.coaguila/Downloads/d912f51470734382d1.pdf>

5. Delgado, E. (2009), Setiembre 22. Ofrecen “Cajita Mágica” para robarse la energía. Listindiario.com. obtenido el 09 abril desde <http://www.afinidadelectrica.com.ar/articulo.php?IdArticulo=214>
6. Gerencia Comercial - Edelnor S.A.A. (2011) Memoria Anual e Informe de Sostenibilidad 2011. Obtenido el 08 de abril 2014 desde <http://www.edelnor.com.pe/Edelnor/ContenidoWeb/html/edelnor31918.html>
7. Grupo EPASA (2014) Editora: Panamá América - PanamáAmerica.com.pa - Redacción de Economía. obtenido el 08 de abril de 2014 desde <http://panamaamerica.com.pa/content/unos-20-millones-pierden-las-empresas-por-robo-de-energ%C3%ADa>
8. Guayasamín Calderón, Eduardo Gabriel. Estudio para el control y reducción de pérdidas de energía eléctrica en un primario de la subestación Barrionuevo perteneciente a la Empresa Eléctrica Quito S.A.[Tesis Previa a la Obtención del Título de Ingeniero Eléctrico]. Carrera de Ingeniería Eléctrica. Escuela Politécnica Nacional. Quito, Junio 2007 disponible en <http://bibdigital.epn.edu.ec/bitstream/15000/340/1/CD-0766.pdf>
9. Gutierrez Marco (2007). Retos en la Gestión de Proyectos de Data Mining. Universidad Politecnica de Madrid Facultad de informática. Departamento de lenguajes y sistemas Informáticos e ingeniería del software. Obtenido el 01 de mayo 2014 desde http://www.dlsiis.fi.upm.es/docto_lsiis/Trabajos20062007/Gutierrez.pdf
10. Camargo H. y Silva M. (2010). Dos Caminos en la búsqueda de patrones por medio de Minería de Datos: SEMMA y CRISP. Artículo de tecnología. Vol. 9 Nro1. Obtenido el 25 de junio desde http://www.uelbosque.edu.co/sites/default/files/publicaciones/revistas/revista_tecnologia/volumen9_numero1/dos_caminos9-1.pdf

11. Herrera, M (2004) Análisis de pérdidas de energía en el sector de distribución eléctrica. Afinidad eléctrica [Revista en línea]. IEEE Sección panamá, El Noticieero, 32. Obtenido el 08 de abril de 2014 desde <http://www.afinidadelctrica.com.ar/articulo.php?IdArticulo=102>
12. KDNUGGETS (2012). Encuesta: ¿Cuál es la Metodología más usada para Minería de datos. Comunidad de Minero de datos- EEUU obtenido el 01 de mayo desde www.kdnuggets.com/polls/2007/data_mining_methodology.htm
13. KDNUGGETS (2012). Encuesta: ¿Cuáles son los campos de aplicación más utilizados en Minería de Datos? Comunidad de Minero de datos- EEUU obtenido el 01 de mayo desde <http://www.kdnuggets.com/2012/12/poll-results-where-did-you-apply-analytics-data-mining.html>
14. Ley Orgánica del Sistema y Servicio Eléctrico (LOSSE). (2011). Gaceta oficial de la República Bolivariana de Venezuela. Obtenido el 16 de abril del 2014 desde <http://www.mppee.gob.ve/uploads/86/69/86694ac1049c41dabdf2623e15611b4/LEY-ORGANICA-DEL-SISTEMA-Y-SERVICIO-ELCTRICO.pdf>
15. Lima Pérez, Liliana (2011). Estrategia Inteligente eficiente para la planificación de las Inspecciones de suministros de las Empresas del Servicio Eléctrico. Tesis Doctorado. Dirección de Investigación y Postgrado Barquisimeto. Doctorado en Ciencias de la Ingeniería, Mención Productividad, Universidad Nacional Experimental Politécnica “Antonio José de Sucre”. República Bolivariana de Venezuela. Obtenido el 08 de abril de 2014 desde <http://es.scribd.com/doc/56008363/Perdidas-electricas-no-tecnicas>
16. Matich, Damian (2001). *Redes Neuronales: Conceptos básicos y aplicaciones*. Obtenido 01 de mayo del 2014, desde http://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientado_ra1/monograis/matich-redesneuronales.pdf

17. Massachussets (2014). The Technology Review Ten. Publicado el 10 de octubre el 2001. Obtenido el 03 de Abril del 2014 desde www.techreview.com
18. Moine Juan, Haedo Ana, Gordillo Silvia (2011). Estudio comparativo de metodologías para minería de datos. Grupo de investigación en Minería de Datos, UTN Rosario Facultad de Ciencias Exactas, Universidad Nacional de Buenos Aires Facultad de Informática, Universidad Nacional de La Plata. Artículo obtenido el 01 de mayo 2014 desde <http://sedici.unlp.edu.ar/handle/10915/20034>
19. Moreno Gustavo (2007). Poder Judicial, Mentoría Bussiness Intelligence. Informe de Planificación y Análisis de requerimientos obtenido el 09 de Junio de 2014 desde http://pmsj.org.pe/memoria2010/MEM_2010_029_entreg2sdb.pdf
20. MR Consultores (2006). Pérdidas No Técnicas. Artículo obtenido el 30 de mayo del 2014 desde <http://www.afinidadelectrica.com.ar/articulo.php?IdArticulo=215>
21. Revovetec (2009). Madrid. Empresa de capacitación en tecnologías, equipos de centrales eléctricas. Obtenida el 10 de abril del 2014 desde <http://www.cicloscombinados.com/sistemaaltatension.html#1. Generador o alternador eléctrico>.
22. Rios Andrés, Uribe Kevin. (2013). Minería de Datos Aplicada a la Detección de Clientes con Alta Probabilidad de Fraudes en Sistemas de Distribución. Proyecto de Grado para optar por el título de ingeniero Electricista. Facultad de Ingenierías. Programa de Ingeniería Eléctrica. Pereira. Obtenido el 10 marzo del 2014 desde <http://repositorio.utp.edu.co/dspace/bitstream/11059/3856/1/006312R586.pdf>
23. Rodríguez Alicia, Castro Sebastián (2012). Detección de Consumos Anómalos de energía eléctrica utilizando nuevas características.

Proyecto CSIC Sector Productivo Modalidad I – UTE. Obtenido el 10 de marzo del 2014 desde https://eva.fing.edu.uy/pluginfile.php/67612/mod_page/content/2/rodriguez_castro-consumos.pdf

24. Santamaría Lema, Fernando Ramiro. Determinación de pérdidas de la red subterránea del alimentador 12 de noviembre de la subestación Atocha y Loreto, de la E.E.A.S.A.[Tesis Previa a la Obtención del Título de Ingeniero Eléctrico]. Facultad de Ingeniería Eléctrica y Electrónica. Escuela Politécnica Nacional. Quito, Septiembre 2008, disponible en <http://bibdigital.epn.edu.ec/bitstream/15000/801/1/CD-1699%282008-10-07-10-55-44%29.pdf>
25. Santamaría Wilfredo (2010). Modelo de detección de fraude basado en el descubrimiento simbólico de reglas de clasificación extraídas de una Red Neuronal. Tesis de grado para optar por el Título de Magister en Ingeniería de Sistemas y Computación, Universidad Nacional de Colombia. Departamento de Ingeniería de Sistemas e Industrial. Bogotá D.C. obtenido el 10 de marzo del 2014 desde <http://www.bdigital.unal.edu.co/3086/1/299742.2010.pdf>
26. Sabino Carlos (1996). El proceso de investigación. Editorial LUMEN/HVMANITAS Viamonte 1674 Buenos Aires. Argentina. Obtenido el 10 de mayo del 2014 desde <http://hugoperezidiart.com.ar/tallerdetesis-pdf/55-sabino-pp1-92.pdf>

ANEXOS

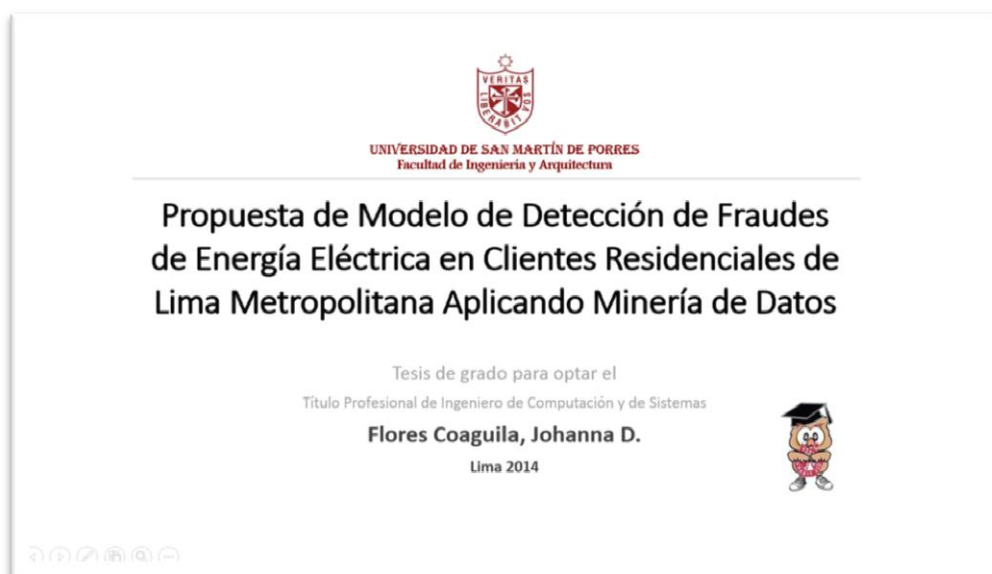


Figura 87: Sustentación Tesis Pag. 01

Elaboración: La autora

Agenda

Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos

UNIVERSIDAD DE SAN MARTÍN DE PORRES

- A. El Problema
- B. Objetivos
- C. Marco Teórico
- D. Metodología
- E. Desarrollo del Modelo
- F. Pruebas y Resultados
- G. Conclusiones
- H. Recomendaciones

2

Figura 88: Sustentación Tesis Pág. 02

Elaboración: La autora

A. El Problema

Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos

UNIVERSIDAD DE SAN MARTÍN DE PORRES

Las empresas de distribución eléctricas no cuentan con un modelo automatizado que apoye en la detección de fraudes de energía eléctrica de clientes residenciales de Lima Metropolitana debido que las modalidades de hurto cambian constantemente, la demanda aumenta y se necesita tratar mayor volúmenes de datos .

¿De qué modo se puede detectar clientes fraudulentos para minimizar las pérdidas de energía eléctrica?

A. El Problema B. Objetivos C. Marco Teórico D. Desarrollo del Modelo E. Pruebas y Resultados F. Conclusiones

3

Figura 89: Sustentación Tesis Pág. 03

Elaboración: La autora

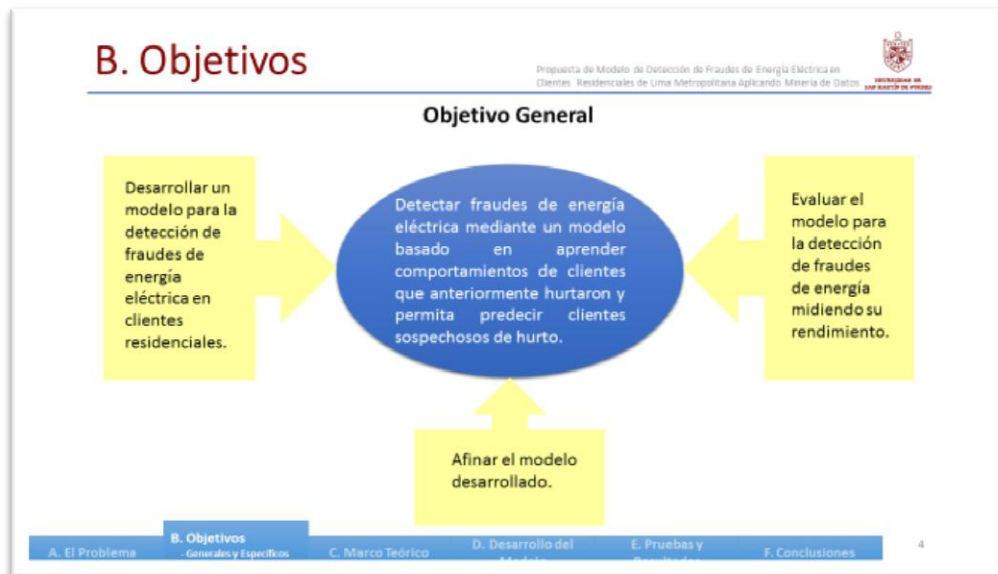



Figura 90: Sustentación Tesis Pág. 04

Elaboración: La autora



Figura 91: Sustentación Tesis Pág. 05

Elaboración: La autora



 Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos

C. Marco Teórico

Red Neuronal Artificial

- Es un sistema de procesamiento de información que tiene propiedades inspiradas en las redes neuronales biológicas.
- ✓ El procesamiento ocurre en las neuronas.
- ✓ Las señales son transferidas a través de enlaces de conexión.
- ✓ Cada conexión tiene un peso asociado representando la sinapsis.
- ✓ Cada neurona aplica una función de activación a su entrada de red para determinar su salida.

Neurona Artificial (1)




$X1 = \text{Entrada 1}$ $wx = (X1 * W1) + (X2 * W2)$
 $X2 = \text{Entrada 2}$ $F(wx) = \text{Tangente Hiperbolica de } wx$
 $W1 = \text{Peso 1}$ $Y1 = \text{Salida 1}$
 $W2 = \text{Peso 2}$

Fuente: Kasperu 2010
Fuente: Matich (2001)

A. El Problema
B. Objetivos
C. Marco Teórico
D. Desarrollo del Modelo
E. Pruebas y Resultados
F. Conclusiones
6

Figura 92: Sustentación Tesis Pág. 06


Elaboración: La autora


 Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos

C. Marco Teórico

Metodología CRISP - DM

- Metodología de procesos de DM que describe los enfoques comunes que utilizan los expertos en minería de datos.
- Se divide en fases, tareas generales, tareas específicas y resultado.
- Ayuda a planear y a administrar proyectos.
- Sin propietario.
- Alienta la inter – operatividad de herramientas.
- Enfocado al negocio y al análisis técnico.



Fuente: Rodríguez, Álvarez, Mesa y González (2010)

A. El Problema
B. Objetivos
C. Marco Teórico
D. Desarrollo del Modelo
E. Pruebas y Resultados
F. Conclusiones
7

Figura 93: Sustentación Tesis Pág. 07

Elaboración: La autora

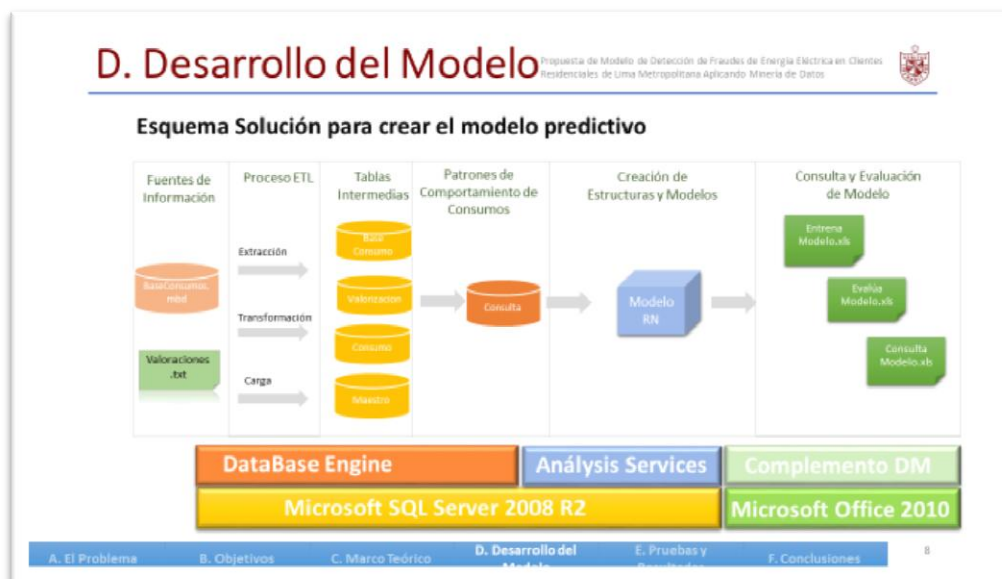


Figura 94: Sustentación Tesis Pág. 08

Elaboración: La autora

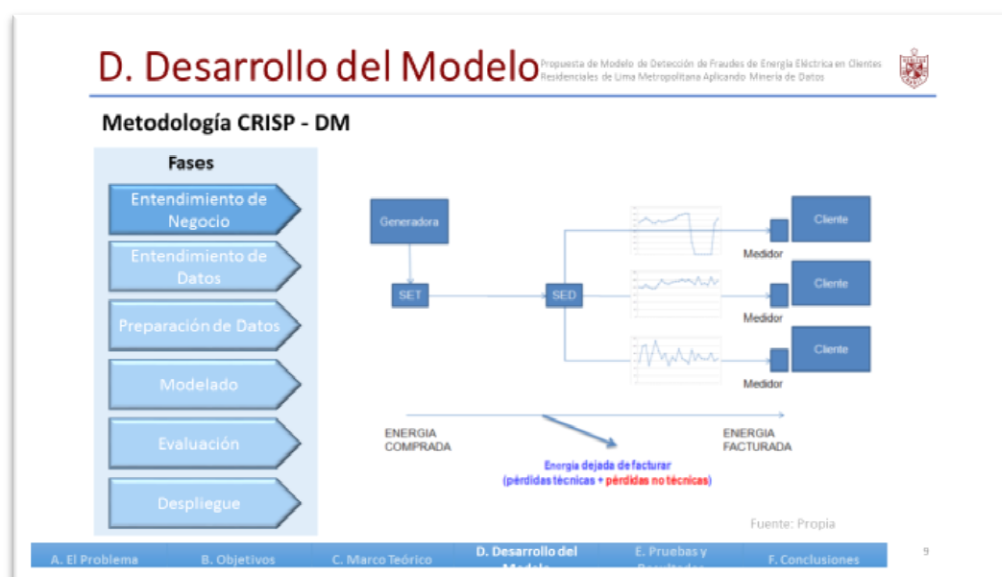


Figura 95: Sustentación Tesis Pág. 09

Elaboración: La autora

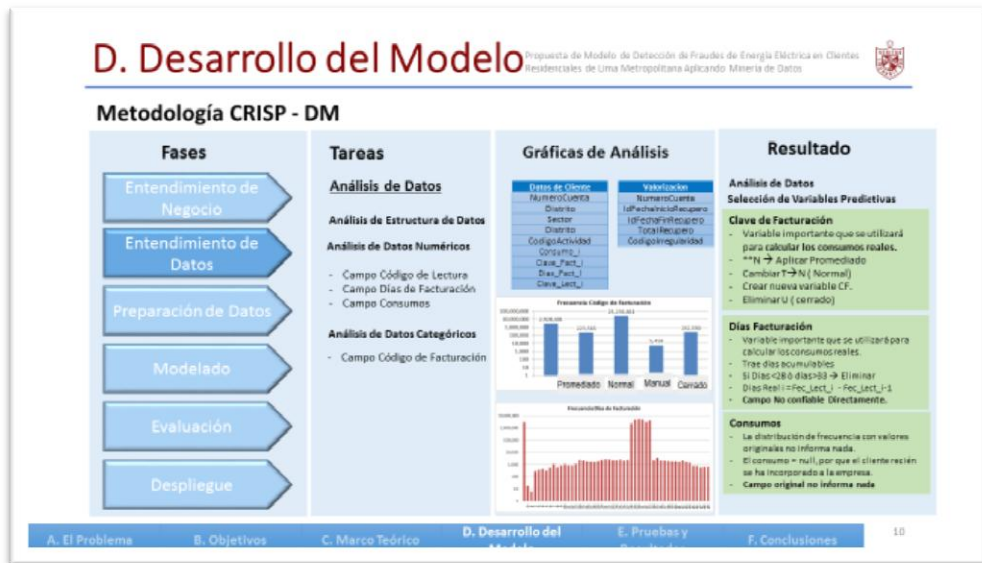


Figura 96: Sustentación Tesis Pág. 10

Elaboración: La autora

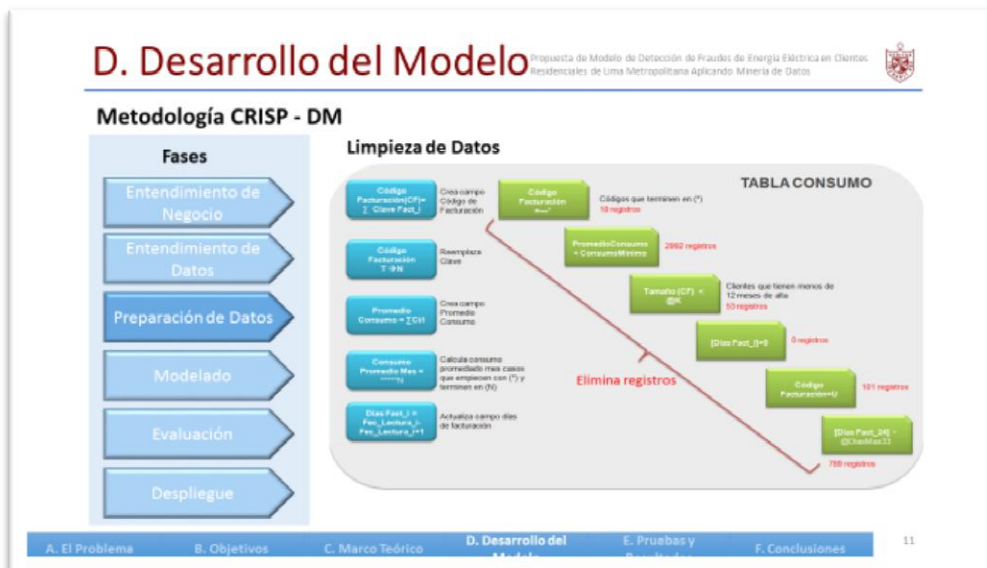


Figura 97: Sustentación Tesis Pág. 11

Elaboración: La autora

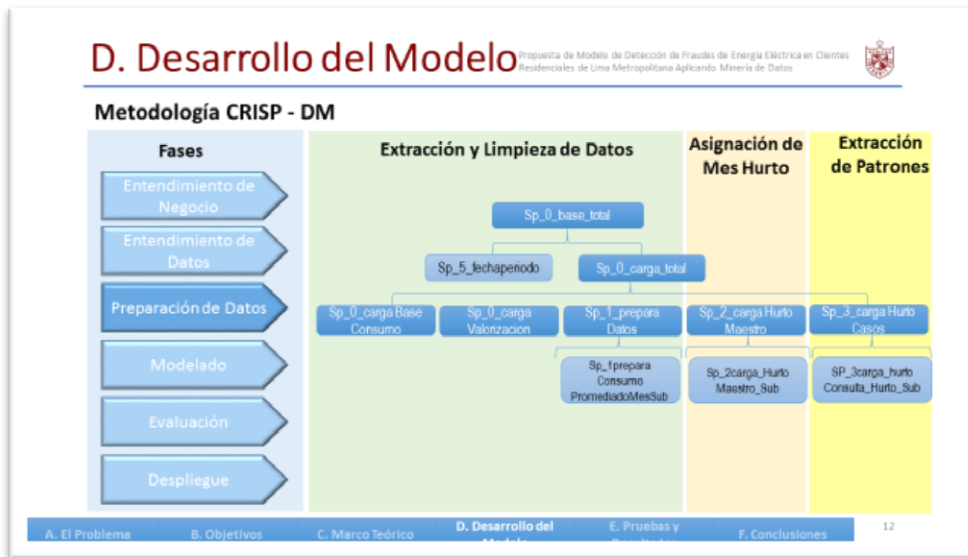


Figura 98: Sustentación Tesis Pág. 12

Elaboración: La autora

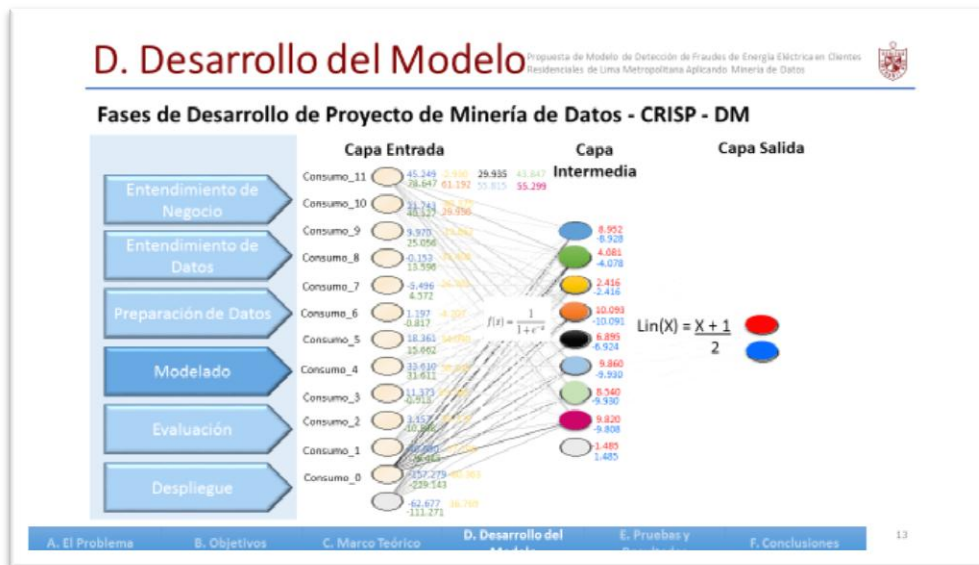


Figura 99: Sustentación Tesis Pág. 13

Elaboración: La autora

D. Desarrollo del Modelo

Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos



Fases de Desarrollo de Proyecto de Minería de Datos - CRISP - DM

Entendimiento de Negocio

$$ON1 = \sum (\text{Consumo}_{11} * 45.246 + \text{Consumo}_{10} * 21.743 + \text{Consumo}_9 * 9.970 + \text{Consumo}_8 * 0.153 + \text{Consumo}_7 * 5.496 + \text{Consumo}_6 * 1.197 + \text{Consumo}_5 * 18.361 + \text{Consumo}_4 * 53.610 + \text{Consumo}_3 * 11.573 + \text{Consumo}_2 * 3.157 + \text{Consumo}_1 * 40.880 + \text{Consumo}_0 * 157.279 - 62.677)$$

Entendimiento de Datos

$$N1 = f \text{ sig } (ON1)$$

$$N11 = f \text{ sig } (ON11)$$

Preparación de Datos

Calcular el Resultado

$$O1 = f \text{ sig } (ON1) * 8.952 + f \text{ sig } (ON2) * 4.081 + f \text{ sig } (ON3) * 2.416 + f \text{ sig } (ON4) * 10.093 + f \text{ sig } (ON5) * 6.895 + f \text{ sig } (ON6) * 9.860 + f \text{ sig } (ON7) * 8.540 + f \text{ sig } (ON7) * 9.820 - 1.485$$

Modelado

$$S = f \text{ lin } (O1)$$

Donde: $Lin(X) = \frac{X + 1}{2}$

Evaluación

Formuladocx

Despliegue

Figura 100: Sustentación Tesis Pág. 14

Elaboración: La autora

D. Desarrollo del Modelo

Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos



Fases de Desarrollo de Proyecto de Minería de Datos - CRISP - DM

Entendimiento de Negocio

Entendimiento de Datos

Resumen del modelo: RedNeuronalPrueba? RedNeuronalPrueba?	
Totales de correctas:	86,99 % 748794
Totales de incorrectas:	13,01 % 111589
Resultados como porcentajes para el modelo "RedNeuronalPrueba?"	
si	86,20 % 44,97 %
no	10,71 % 55,03 %
Correcta	86,20 % 55,03 %
Incorrecta	10,71 % 44,97 %
Resultados como recuentos para el modelo "RedNeuronalPrueba?"	
si	754195 25966
no	85623 83770
Correcta	754195 83770
Incorrecta	85623 25966

Preparación de Datos

Modelado

Evaluación

Despliegue

Figura 101: Sustentación Tesis Pág. 15

Elaboración: La autora

D. Desarrollo del Modelo

Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos



Fases de Desarrollo de Proyecto de Minería de Datos - CRISP - DM

id	NumeroCuenta	Prediccion	Prob. SI	Prob. No
1	3	SI	0.71113821	0.28874182
1	88	no	0.325875254	0.674024746
2	4	SI	0.810675644	0.189324356
2	88	no	0.266376107	0.633523923
3	5	no	0.396891273	0.603008727
3	88	no	0.37052324	0.62947676
4	7	SI	0.511634493	0.488365507
4	88	no	0.343390887	0.656609113
5	8	no	0.462550423	0.537449577
5	88	no	0.382059881	0.617940119
6	5	no	0.487232888	0.512767112
6	88	no	0.37752397	0.62247603
7	11	SI	0.558491331	0.441508669
7	88	no	0.347989266	0.652010734
8	12	no	0.438905477	0.561094523
8	88	no	0.385404453	0.614595547
9	13	no	0.495461903	0.504538097

Figura 102: Sustentación Tesis Pág. 16

Elaboración: La autora

E. Pruebas y Resultados

Propuesta de Modelo de Detección de Fraudes de Energía Eléctrica en Clientes Residenciales de Lima Metropolitana Aplicando Minería de Datos



Modelo	Casos Hurto	Casos No Hurto	Técnica	Evalúa	Sensibilidad	Especificidad	Valor de éxito
RN_01 200910	10000	20000	Consumo Promedio Diario Promedio mayor a 11 KW	100000 Aleatorio	23.08%	97.34%	97.3%
RN_02 200910	20000	30000	Consumo Promedio Diario Promedio mayor a 11 KW	857563 Todos	48.89%	92.72%	89.47%
RN_03 200910	20000	20000	Consumo Promedio Diario Promedio mayor a 11 KW	857563 Todos	55.03%	89.29%	86.99%
RN_04 200910	20000	20000	Consumo Promedio Diario Normalizado Promedio mayor a 11 KW	857563 Todos	63.64%	26.33%	27.97%
RN_05 200910	20000	30000	Consumo Promedio Diario Normalizado Promedio mayor a 11 KW	857563 Todos	67.88%	40.42%	66.67%

A. El Problema B. Objetivos C. Marco Teórico D. Desarrollo del Modelo E. Pruebas y Resultados F. Conclusiones 17

Figura 103: Sustentación Tesis Pág. 17

Elaboración: La autora

F. Conclusiones



- Se logró desarrollar modelo con un **55.03%** sensibilidad y **de 89%** de especificidad, siendo su **valor de éxito el 86.99%** mientras que el **proceso tradicional** según la memorial anual publicada por Edelnor en el 2011 se realizaron 193,516 inspecciones se logra regularizar el **17419** consumos no registrados equivalente al **9% de éxito**.
- El modelo desarrollado se obtuvo de datos balanceados **20000 (50%)** casos de hurto y **20000 (50%)** casos de no hurto.
- La preparación de los datos tomó el 60% de esfuerzo.
- La normalización como técnica para afinar el modelo mejora la sensibilidad del modelo pero no es bueno aprendiendo los casos de no hurto.

Figura 104: Sustentación Tesis Pág. 18

Elaboración: La autora

Recomendaciones



- El modelo de detección de fraude se realizó sobre energía eléctrica pero también se sugiere realizar el mismo proceso para detectar perdidas en otros servicios agua, gas, etc.
- Se sugiere hacer pruebas de campo y validar con los técnicos el resultado de la detección.
- Para detectar clientes fraudulentos de energía eléctrica en la actualidad se sugiere volver a generar un modelo con los procedimientos establecidos en la tesis pero con información actualizada de un periodo anterior hacia atrás 25 meses del periodo que se desea predecir.

Figura 105: Sustentación Tesis Pág. 19

Elaboración: La autora