Modelo de caracterización del perfil del cliente adaptado al sector de la construcción, aplicando minería de datos: caso de estudio en empresa constructora

Characterization model of customer profile adapted to building sector with data mining: case study building company

Carlos Alberto Erazo Cardona
carlos.erazo02@usc.edu.co

Yana Saint-Priest Velásquez, M.Sc
yana.saint-priest00@usc.edu.co

Universidad Santiago de Cali, Facultad de Ingeniería, Programa de Maestría en Informática
Grupo de investigación GIEIAM

Resumen

El proceso de edificación de viviendas en las constructoras se ve afectado por el desistimiento de las negociaciones antes de la entrega material del bien inmueble, desconociéndose las razones que ocasionan esta situación. En el sector de la construcción no hay un modelo que permita la caracterización del perfil del cliente que se vincula al proceso de comercialización de viviendas que contribuya a mitigar el riesgo de caída de los acuerdos de negocio. A medida que los diversos sectores productivos empiezan a aplicar minería de datos en la resolución de sus problemas, surgen adaptaciones de las metodologías tradicionales, que solucionan particularidades de dichos sectores. El presente artículo propone un modelo para la caracterización del perfil del cliente, adaptado al sector de la construcción, aplicando minería de datos, que incluye los resultados obtenidos en un caso de estudio. Para el desarrollo se utilizó la metodología de descubrimiento de conocimiento en bases de datos (KDD) con una adaptación o contribución encontrada en la literatura, que permitió identificar características propias del gremio de la construcción en el planteamiento, diseño, técnicas de minería, evaluación de patrones y la validación, que posibilitaron generalizar el modelo. Con el caso de estudio se pudo plantear un diseño para la bodega de datos basado en el modelo de datos multidimensional utilizando el esquema en estrella. Se comprobó que las técnicas de clasificación, regresión y agrupamiento son las recomendadas para resolver el problema de investigación y finalmente se presentó la propuesta para la evaluación y validación de las técnicas de minería de datos. La generalización del modelo de caracterización del perfil del cliente, ratificó la tendencia y requisito de los sectores productivos de contar con modelos de minería de datos particulares a sus necesidades.

Palabras Clave: Minería de datos, Bases de datos, Modelo de Minería de datos

Abstract

Building houses process in construction companies is affected by fallen negotiations before delivery of the living place, unknowing the reasons for this situation. In building sector there is no a model that allows the characterization of customer profile linked to the commercialization process that would contribute to mitigate the risk to business agreement. As several productive sectors begin to apply data mining, adaptations of traditional methodologies for data mining arise, which solve particularities of these sectors. This paper proposes a characterization model of customer profile adapted to building sector with data mining that includes the results obtained in case study in a building company. The Knowledge Discovery in Databases (KDD) methodology was followed with an adaptation or contribution found in literature that identified own characteristics of building industry in the approach, design, mining techniques, pattern evaluation and validation that allowed generalizing the model. With case study, a design for the data warehouse based on the multidimensional data model using the star schema was proposed. It was proved that classification, regression and clustering techniques are recommended to solve the problem and finally the proposal for data mining techniques evaluation and validation was presented. The generalization of the characterization model of customer profile, confirmed the tendency and requirement of the productive sectors to have data mining models specific to their needs.

Keywords: Data Mining, Data Base, Data Mining Model

1. INTRODUCCIÓN

El desistimiento de las negociaciones en el proceso de comercialización de viviendas en las constructoras afecta notablemente el desarrollo de los proyectos, debido a que el inicio de obras se da sólo si las ventas han alcanzado el punto de equilibrio. Según el informe: Colombia Construcción en Cifras (Cámara Colombiana de la Construcción CAMACOL, 2018), las obras paralizadas durante el año 2018, respecto a las obras activas representan un 28,58%. Un proyecto clasificado como inactivo significa que su avance de construcción es cero durante el año evaluado y aplica para los proyectos que no han alcanzado el punto de equilibrio.

Una causa relevante que incide en los desistimientos tiene que ver con el hecho que no se analicen los datos transaccionales del negocio. Los procesos de comercialización de viviendas se encuentran hoy en día bien definidos y muchas tareas de este proceso están automatizadas con software, incluso especializado como CRM, pero se está desaprovechando el potencial de los datos que resultan después del procesamiento cotidiano que queda almacenado en bases de datos. Otra perspectiva causal del problema, se relaciona con la administración de la relación con el cliente, recordando que la satisfacción del cliente es el referente del éxito del proceso de comercialización de viviendas. En este contexto las familias no sólo buscan la compra de un lugar de residencia, sino que también quieren ver materializado el cumplimiento de un sueño. Como parte del mejoramiento continuo del proceso comercial, las constructoras deben acercarse al conocimiento inmediato del cliente objetivo.

La minería de datos es un concepto que surge como respuesta a la necesidad del análisis de grandes volúmenes de información. De acuerdo con (Han & Kamber, 2011), la minería de datos puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias, al examinar gran cantidad de datos. El aporte que brindan las técnicas de minería de datos, consiste en generar conocimiento nuevo a partir de datos que a simple vista parecen irrelevantes o que ya han cumplido su función tras el procesamiento. En un mundo de negocios altamente automatizado, las necesidades de almacenamiento están creciendo a un ritmo elevado; cada vez se requiere más espacio para los datos procesados y por ende las operaciones de cómputo también se incrementan (Lee, 2017). De esta realidad nacen interrogantes sobre qué hacer con tanta cantidad de datos y si el único destino debe ser un repositorio de datos.

La minería de datos es una disciplina reciente, en la cual interactúan diversas áreas, especialmente la estadística, el aprendizaje automático y la gestión de bases de datos para intentar encontrar patrones en grandes volúmenes de datos (Azzalini & Scarpa, 2012). De acuerdo a lo anterior, la minería de datos es una disciplina que requiere una comprensión tanto estadística como computacional para poder abordar problemas de investigación.

Existen numerosas técnicas de minería de datos aplicables a los problemas de investigación, (Han & Kamber, 2011), enfatizan en la técnica de clasificación como una forma de análisis de datos que extrae modelos que describen clases de datos relevantes. Estos modelos se reconocen como clasificadores y predicen las etiquetas de clase categóricas. Por su parte, clustering es el proceso de agrupar un conjunto de datos en múltiples grupos de tal manera que los objetos dentro de un grupo tengan una gran similitud, pero son muy diferentes a los objetos en otros grupos. En contraposición con la clasificación, la etiqueta de clase es desconocida y el objetivo de la técnica es hallar dicha etiqueta.

Cuando el tamaño del conjunto de datos es muy grande, la computación en paralelo ofrece una manera eficiente de encontrar conjuntos de elementos frecuentes en menor tiempo (Vasoya & Koli, 2016). Otro avance en la minería de datos a través de computación paralela se basa en el modelo MapReduce (Yong, Jiecai, & Xueqing, 2018).

El campo de acción de la minería de datos es diverso, como lo relacionan (Rygielski & Jyun-Cheng, 2002), existen numerosas áreas donde la minería de datos se puede aplicar; esto es, prácticamente en todas las actividades humanas que generen datos.

En el sector de la educación, (Al-Twijri & Noamanb, 2015), plantean un modelo de minería de datos adaptado a instituciones de educación superior. En él, los autores exponen la necesidad de utilizar los datos educativos, que hoy en día presentan un crecimiento explosivo, para mejorar la calidad de las decisiones de gestión. El modelo sugerido asiste en el proceso de toma de decisiones en los niveles estratégicos de instituciones superiores, así como también regula las disciplinas de admisión de los estudiantes. El sector de logística y distribución, siempre ha enfrentado grandes retos para

hacer llegar los productos o servicios finales a manos del consumidor. (Listanti, 2014), propone una gestión proactiva del rendimiento de la cadena de suministro a través de analítica predictiva.

Ahora bien, la administración de la relación con el cliente es una herramienta que toma protagonismo en las organizaciones, teniendo en cuenta que los clientes representan las ventas y en consecuencia los ingresos. (Dariush, Hojjatollah, Hajigol, & Parirooy, 2016), investigaron la comprensión del comportamiento del cliente bancario utilizando minería de datos, logrando obtener valiosa información respecto a oportunidades específicas de inversión.

Un marco de trabajo para la minería de datos en la administración de la relación con el cliente (CRM) es propuesto por (Bahari & Elayidom, 2015), en él los autores plantean que en la primera fase es necesario comprender los objetivos comerciales y los requisitos del dominio del problema. Un estudio cercano y la gestión de las relaciones con los clientes y sus interacciones ayudarán a identificar, atraer y retener clientes en el dominio.

La minería de datos también ha incursionado en el sector de la construcción. En la ciudad de Poznan en Polonia, (Gajzler, 2016), realizó una investigación sobre descubrimiento de conocimiento, a través de dos casos de estudio. El primero de ellos más técnico, tiene que ver con la selección de un mortero adhesivo para baldosas de cerámica y el segundo la definición de criterios de selección y compra por parte de los clientes, basados en la ubicación de las unidades de apartamentos dentro de la ciudad. En Turquía, se desarrolló un caso de estudio relacionado con un problema en la industria de la construcción asociado con la productividad y la calidad de los equipos (grupos de personas) que producen baldosas de cerámica. (Kaya, Keles, & Oral, 2014), realizaron un estudio utilizando minería de datos que pudiera describir la baja calidad y el alto costo del producto final.

Los anteriores casos de estudio ilustran cómo el sector de la construcción empieza a adaptar la minería de datos a sus características particulares. Adecuar las técnicas de minería de datos a contextos específicos permite un mejor planteamiento de los objetivos y por ende mejores resultados. Existe la oportunidad de poder generar un modelo que incluya varias características propias del sector de la construcción; entre ellas, el enfoque hacia el cliente en el proceso de comercialización de viviendas, los tipos de bases de datos, entidades y relaciones que se asocian en este contexto, todo entrelazado y soportado con la literatura sobre minería de datos y los resultados del caso de estudio, lo que se convirtió en el objetivo de esta investigación.

El resultado de la presente investigación consistió en el diseño de un modelo de caracterización del perfil del cliente, adaptado al sector de la construcción aplicando minería de datos, que incluyó el trabajo y resultados de un caso de estudio en una empresa constructora. El artículo es presentado en tres secciones: metodología, resultados y por último, la sección de conclusiones.

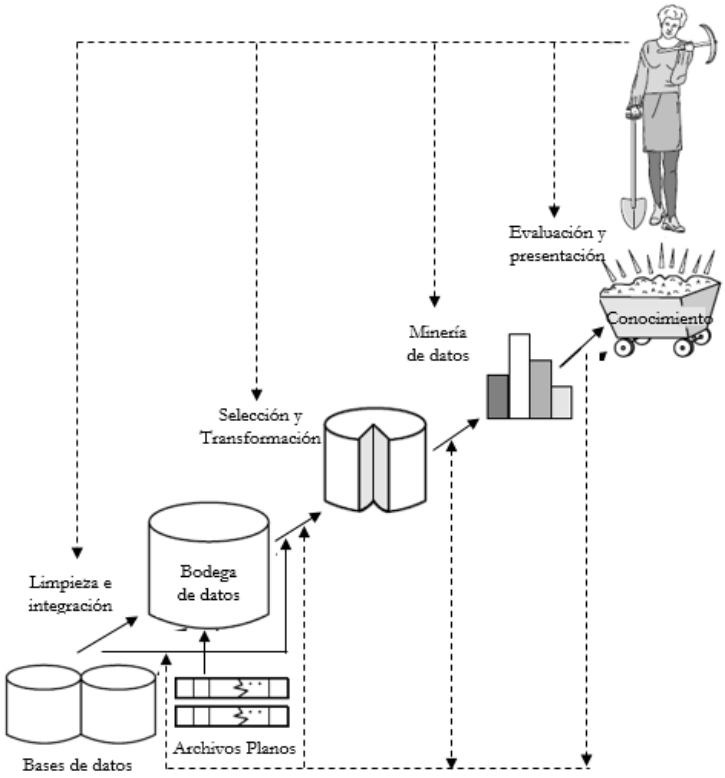
2. METODOLOGÍA

El problema de investigación fue abordado inicialmente con el análisis y contextualización del mismo en el sector de la construcción y en el caso de estudio. Para validar lo que los estudios del gremio evidenciaron, se realizó un diagnóstico en el caso de estudio: Constructora Moreno Tafurt S.A. Ésta es una empresa dedicada a la construcción y venta de vivienda en el Valle del Cauca, enfocada principalmente en proyectos de interés social y afiliada a CAMACOL. La compañía, fue constituida en el año 2003 y cuenta con 7 proyectos de construcción de viviendas en curso.

La información del proceso de comercialización de las viviendas de todos los proyectos es registrada en el CRM: Smart-Home, de acuerdo al contrato firmado entre las partes (GAIA TELCOM Sucursal Colombia y Consorcio Moreno Tafurt S.A., 2014). Antes de interactuar con la base de datos, se realizaron entrevistas con la líder del proceso comercial, con 4 ejecutivas pertenecientes a fuerza de ventas y se asistió a las salas de exhibición de los proyectos para conocer en detalle el contacto inicial del cliente con la constructora.

Para el descubrimiento de conocimiento a través de técnicas de minería de datos, se utilizó la metodología de descubrimiento de conocimiento en bases de datos (KDD), ilustrada en la Figura 1, citada y adaptada por (Han & Kamber, 2011).

Figura 1: Proceso de descubrimiento de conocimiento



Fuente: Adaptada de (Han & Kamber, 2011)

La metodología KDD se seleccionó debido a que ofrece independencia en la implementación, en contraposición con la metodología de muestreo, exploración, modificación, modelado y evaluación SEMMA; que de acuerdo con (Matignon, 2007), está diseñada para contribuir a la usabilidad del software SAS Enterprise Miner. Por otra parte, la metodología de minería de datos para procesos de la industria CRISPDM que también fue analizada, ofrece una visión muy industrial y posee adaptaciones recientes (Steffen, Hajo, Dorothea, & Steffen, 2019) para cumplir con los retos que plantean las nuevas tecnologías y la disponibilidad de diversas fuentes de datos.

2.1. Limpieza e integración

La fuente de datos correspondió a la base de datos del CRM de la compañía. En la limpieza de datos se removió el ruido correspondiente a atributos y entidades no relevantes para el problema de investigación; por citar algunos, se retiraron entidades como: cargo del vendedor, tipos de identificación y medios de publicidad. Se identificaron atributos del tipo booleano que para el diseñador de la base de datos significaban ausencia o presencia de valor y se crearon analogías o significados para otros atributos del tipo lista (1,2,3,4…), que no se encontraban normalizados.

2.2. Diseño del data warehouse y transformación

Para el diseño del data warehouse se seleccionó el modelo de datos multidimensional o cubo de datos y dentro de los esquemas de datos multidimensionales el esquema en estrella resultó el más acertado. La razón por la cual se eligió el modelo multidimensional obedece a la mejora y ajuste de los datos bajo el paradigma de dimensiones y hechos, evitando las típicas consultas SQL sobre las bases de datos que pueden consumir tiempo significativo de procesamiento en el cálculo de JOINS e índices. De otro lado, dentro de los esquemas disponibles en el modelo multidimensional, se utilizó el esquema en estrella, que comparado con el copo de nieve y el esquema constelación, ofrece mayor eficiencia, ya que las dimensiones no se normalizan con otras tablas y esto hace que las consultas no sean complejas como normalmente ocurre en las bases de datos transaccionales.

Las dimensiones seleccionadas fueron: cliente-tiempo-pagos-proyecto y como tabla de hechos se asignaron las ventas. Las métricas pueden apreciarse en la tabla de hechos, particularizadas contra cada dimensión o combinación de 2 o más

de ellas. Las dimensiones propuestas pueden encontrarse comúnmente en los procesos de comercialización de viviendas, ya que hacen parte del proceso general. Por su parte, las métricas son la consecuencia de operaciones aritméticas o de grupo con atributos que representan variables continuas, que pueden ser calculadas a partir de los datos; en este caso, medidas alrededor de las ventas. Los atributos que se seleccionen para integrar cada dimensión pueden variar, dependiendo de las reglas del negocio de cada constructora; esto es, depende de qué tanta cantidad de columnas posea cada entidad en la base de datos.

Una vez establecido el diseño, se procedió a migrar los datos al data warehouse, lo cual puede hacerse mediante la generación de vistas o consultas a la base de datos transaccional o la migración mediante herramientas especializadas o procesos ETL como lo señalan (Abdellah, Rachid, & Belaid, 2016).

2.3. Minería de datos

Se ejecutaron técnicas de minería de datos sobre el conjunto de datos transformados y migrados al data warehouse, concretamente árboles de decisión, clustering, clasificación bayesiana, regresión logística, redes neuronales y reglas de asociación; en búsqueda de patrones relacionados en los datos. Existen diversas técnicas de minería de datos; para el problema de investigación se emplearon técnicas de: clasificación, regresión, asociación y agrupamiento con el enfoque de aprendizaje supervisado.

Existen numerosas herramientas de software para realizar la implementación de las técnicas de minería de datos. (Naika & Samantb, 2016) detallan algunas de estas herramientas, pero pueden encontrarse muchas más, entre ellas: Orange, RapidMiner, Knime, Weka, Oracle Data Mining, Microsoft Analysis Services, Apache Hadoop o incluso desarrollos propios que incluyan la codificación de los algoritmos base de cada técnica.

La selección de la herramienta de software para minería de datos depende en cada caso de estudio de varios factores; entre ellos, las licencias, el tamaño de la entrada (datos) y las facilidades de integración. Para el caso de estudio abordado se seleccionó Microsoft Analysis Services, que es un motor de datos analíticos utilizado en el soporte de decisiones y análisis de negocios y que proporciona modelos de datos semánticos de nivel empresarial y minería de datos (Microsoft, Corporation, 2018). También se tuvo en cuenta que el motor de la base de datos del CRM es Microsoft SQL Server y se inclinó la preferencia gracias a la posibilidad de uso de la licencia comercial y el tamaño de la entrada de datos de alrededor 25.000 negocios de venta de vivienda.

2.4. Evaluación de patrones y validación de los modelos

Una vez ejecutadas las técnicas de minería de datos, se procedió al análisis e interpretación de los resultados, bajo la premisa que cada técnica produce un modelo. Existe un conjunto de datos de entrenamiento y de prueba, éstos se constituyen es una parte importante de la evaluación de los modelos de minería de datos. Normalmente, al dividir un conjunto de datos en uno de entrenamiento y uno de prueba, la mayor parte de los datos (70%) se usan para el entrenamiento y la menor (30%) para las pruebas.

Los resultados de cada modelo fueron revisados en el contexto del problema de investigación; es decir, interpretando con el área comercial qué tan verídicos podrían ser los resultados obtenidos. Los modelos se validaron haciendo uso de gráficos de elevación, que miden la probabilidad de predicción de cada modelo. También se validaron con la matriz de confusión que, de acuerdo con (Adekitan, Abolade, & Shobayo, 2019) evalúa el desempeño de los algoritmos de aprendizaje supervisado mediante la presentación de falsos positivos, verdaderos positivos, falsos negativos y verdaderos negativos. Por último se realizó validación cruzada que como lo relacionan (Kittipong, Sukree, & Pongpun, 2019) es una herramienta para medir la precisión de los modelos e incluye pruebas de clasificación (fallo o éxito) y de probabilidad (elevación, puntuación logarítmica y error cuadrático).

Posterior a la validación de los patrones encontrados, se presentaron los resultados específicos del caso de estudio al área comercial de la Constructora Moreno Tafurt S.A., para ser utilizados en la toma de decisiones que propendan por el mejoramiento continuo del proceso de comercialización de viviendas y por ende de toda la organización. Se explicó y detalló a los grupos de interés, toda la metodología utilizada antes de presentar los resultados.

2.5 Generalización: presentación del modelo de caracterización del perfil del cliente

Se presenta el aporte al gremio de la construcción, gracias a la generalización del modelo de caracterización del perfil del cliente que se vincula a los procesos de comercialización de viviendas, que son comunes en las constructoras. En la sección de Resultados, se puede apreciar el modelo paso a paso con sus respectivas explicaciones.

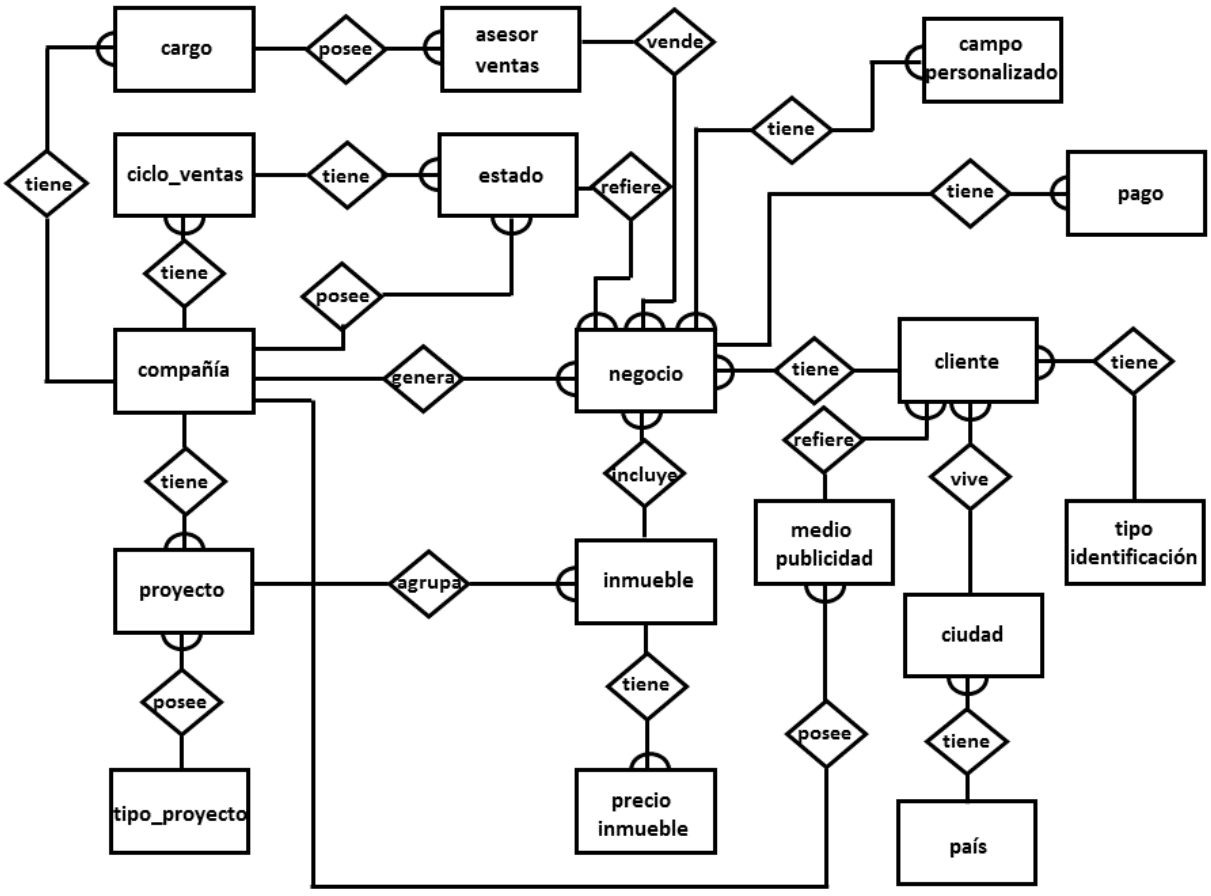
3. RESULTADOS Y DISCUSIÓN

Los resultados de la investigación se dividen en dos partes: primero, un resumen de los resultados correspondientes al caso de estudio y segundo, los que se relacionan con la generalización o presentación del modelo de caracterización del perfil del cliente para empresas constructoras.

3.1. Limpieza e Integración

Como resultado de esta etapa, se pasó de tener un esquema con más de 17 entidades (ver Figura 2), a tener un énfasis en 4 conceptos: el cliente, el negocio, los pagos y el proyecto.

Figura 2: Diagrama entidad relación del CRM del caso de estudio



Fuente: Construcción Propia

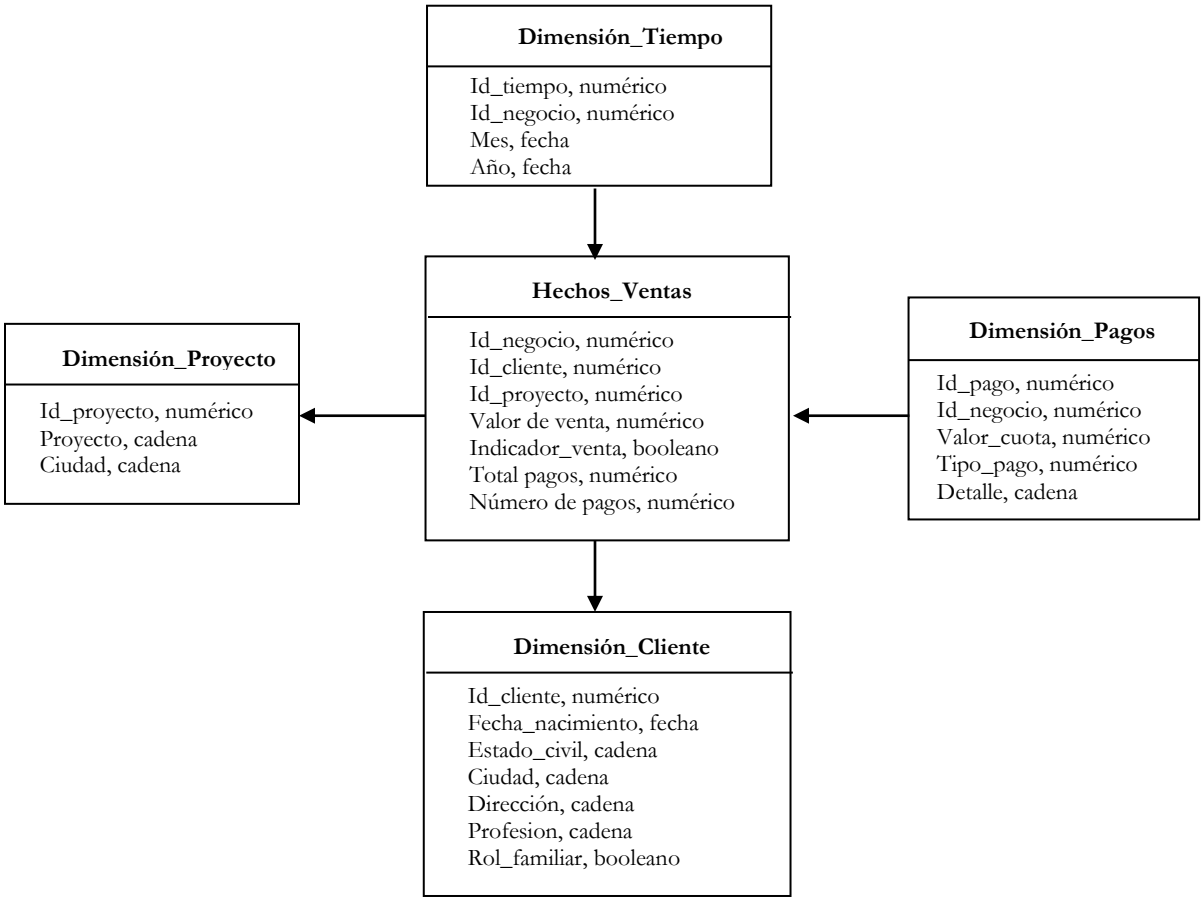
Gracias a la contextualización y comprensión del proceso comercial en el caso de estudio, se estableció que el embudo de ventas empieza con el registro de clientes (visitantes) que hacen el contacto por primera vez. En caso de firmar un acuerdo de pago (a través de un contrato de reserva) se genera un negocio. Finalmente los negocios pueden tener o no pagos registrados. El interés se centró en los negocios de los clientes que tuvieron por lo menos un pago, lo cual aumenta la probabilidad de compra de una vivienda.

La reducción de atributos se realizó conforme al análisis de presencia de valores completos hasta el momento de tener una negociación con pago parcial. De la misma manera, se descartaron los atributos no obligatorios y no fue necesaria la integración de otros repositorios de datos diferentes a la base de datos del CRM.

3.2. Diseño del data warehouse y transformación

Una vez definido en la metodología el uso de cubos multidimensionales basados en el esquema en estrella, se construyó el modelo de datos para el data warehouse ilustrado en la Figura 3. Para poblar la bodega de datos, se hizo uso de conexión directa entre Analysis Services y SQL Server, especificando fácilmente el origen de datos y la conexión al servidor.

Figura 3: Diseño del data warehouse



Fuente: Construcción Propia

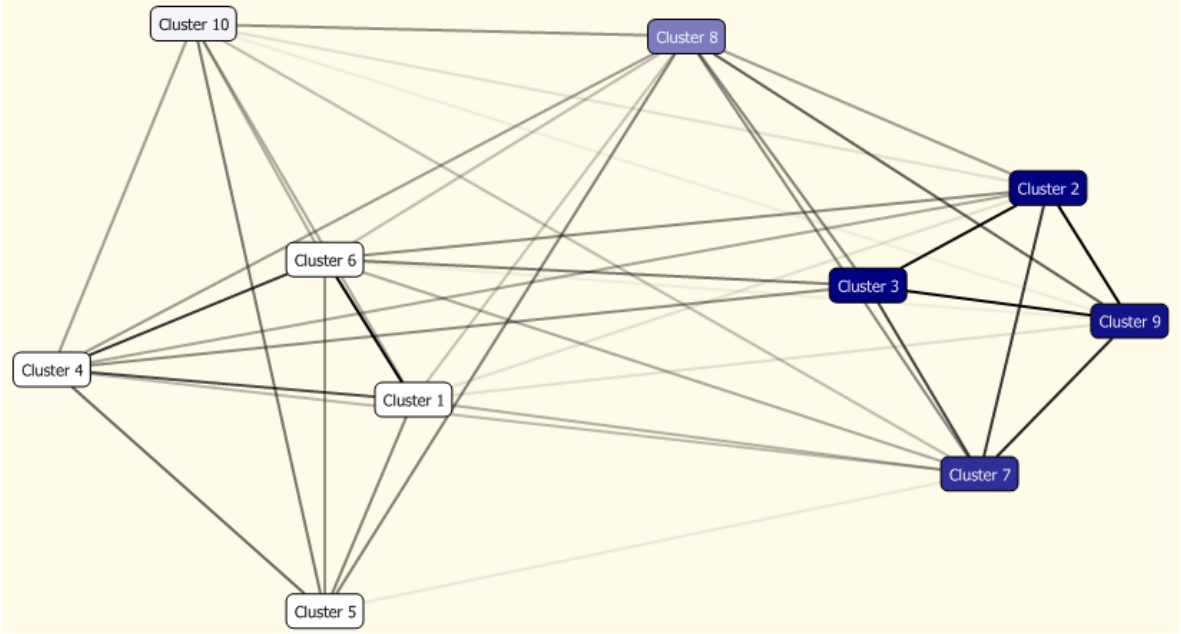
3.3. Minería de datos

La ejecución de las técnicas de minería de datos, permitió conocer resultados diversos. A continuación se presenta un resumen de ellos:

3.3.1. Clustering o agrupamiento

De acuerdo con (Witten, Frank, & Hall, 2011), las técnicas de agrupación se aplican cuando no hay una clase para predecir, pero los casos se dividen en grupos naturales. La Figura 4, muestra la ejecución de la técnica de clustering sobre la entrada de datos. En ésta se observan 10 grupos, los más claros denotan clientes no compradores y los más oscuros se caracterizan como clientes compradores. Cada grupo posee un conjunto de características propias y cercanas entre ellas a la población que alberga.

Figura 4: Resultado de la ejecución de la técnica de clustering sobre el data warehouse



Fuente: Extraído de la herramienta Microsoft Analysis Servicies

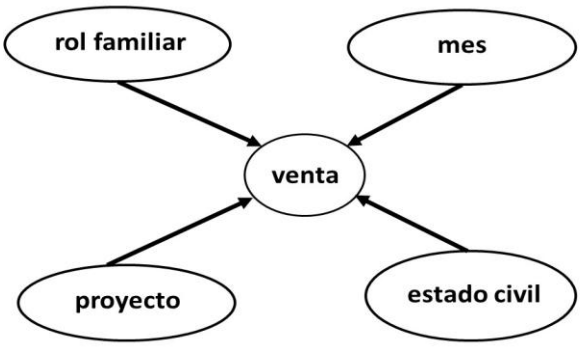
A continuación se detallan dos agrupaciones con las más altas probabilidades:

Clúster 3 - Comprador (99% de probabilidad) → edad promedio: 42 años, estado civil: soltero, valor promedio de compra: $48.269.155, ocupación: empleado, número de cuotas que utiliza para la compra: 6, mes de legalización de la compra: noviembre. Clúster 1 – No comprador (99% de probabilidad) → edad promedio: 30 años, estado civil: soltero, valor promedio de compra $47.362.730, ocupación: independiente, número de cuotas que utiliza para la compra 1, sin mes de legalización de compra.

3.3.2. Clasificación Bayesiana

La clasificación bayesiana ignora los atributos continuos y se enfoca en los de tipo discreto. Para estimar los parámetros sólo necesita una pequeña parte de los datos de entrenamiento (Dutta, Dutta y Raahemi, 2017). Un resultado de la ejecución de la técnica puede apreciarse en la Figura 5, la cual corresponde a una red de dependencia. La técnica también permite conocer los perfiles de cada atributo (valores o estados de los atributos para compradores y no compradores) y la caracterización de acuerdo a la probabilidad de ocurrencia.

Figura 5: Un resultado de la ejecución de la técnica de clasificación bayesiana sobre el data warehouse



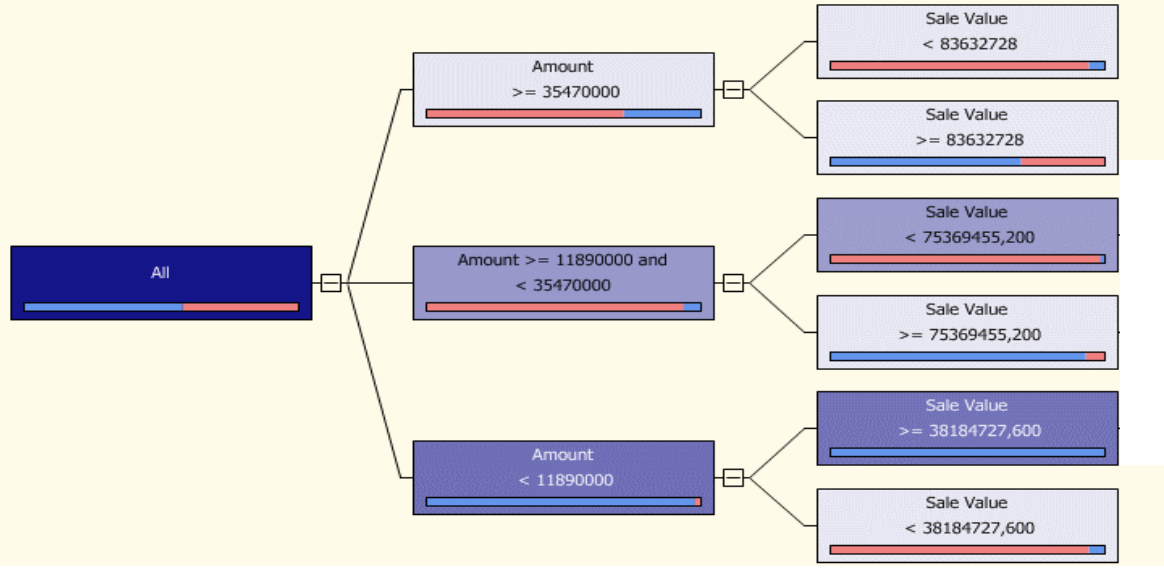
Fuente: Construcción Propia

Un comprador tiene una probabilidad del 90.74% de no ser cabeza de hogar y una probabilidad del 64.59% de ser soltero. Por su parte, si la preferencia del comprador es hacia el proyecto: Urbanización Reserva de Zamorano 604, tiene una probabilidad del 31.41% de ser comprador. Por último el mes en que un cliente legalizará su compra, tiene una probabilidad del 24.89% de ser: noviembre.

3.3.3. Árboles de Decisión

De acuerdo con (Patel, Abbasi, Saeed, & Alam, 2018), un árbol de decisión es un algoritmo de aprendizaje supervisado que proporciona representación visual para la clasificación de un conjunto de datos. Una vez ejecutada dicha técnica, los resultados se inclinaron más hacia las variables continuas, tal y como lo ilustra la Figura 6.

Figura 6: Resultado parcial de la ejecución de la técnica de árboles de decisión sobre el data warehouse



Fuente: Extraído de la herramienta Microsoft Analysis Servicies

Los nodos más oscuros indican mayor concentración de número de casos. En cada nodo, se comparan por histograma, la probabilidad de encontrar compradores y no compradores. Hasta el nivel 3, la ruta más influyente de compradores (mayor población) se compone de clientes cuyos abonos están en el rango de $11.890.000 y $35.470.000 (42.11% de los casos), donde a su vez el 93,27% corresponde a compradores efectivos. En el siguiente nodo de esta rama, hay mayor concentración de compradores si el valor de venta de la vivienda es menor a $75.369.455 (95.07% de los casos del nodo anterior), donde a su vez el 97,74% corresponde a compradores efectivos.

3.3.4. Redes Neuronales

Como lo enfatizan (Dutta et al, 2017), las redes neuronales son algoritmos de aprendizaje automático supervisado o no supervisado, inspirados en las neuronas del sistema cerebral humano La ejecución de la técnica de redes neuronales mostró resultados diversos al poder combinar los posibles estados de los atributos de entrada con el atributo de predicción (venta efectiva). En la Tabla 1, se muestra la puntuación de importancia para los valores del atributo: proyecto de vivienda y muestra cómo están relacionados con el atributo de predicción (compra).

Tabla 1: Ejecución de la técnica de redes neuronales sobre el data warehouse

Proyecto de construcción	Puntuación de importancia (comprador)	Puntuación de importancia (no comprador)
Condominio Monteverde 42	**63.14**	
Chapinero sur VIPA	**24.73**	
Urbanización La Colina	**18.76**	
Urbanización Mi Sueño	**64.13**	
Reserva de Zamorano 314 Etapa 2	**45.71**	
Reserva de Zamorano 604 Etapa 2	**47.83**	
Parques de Versalles		**81.06**
2ª Etapa Monte Verde 42 Condominio		**69.13**
Condominio Reserva Verde 23		**60.03**
2ª Etapa urb. Chapinero Sur VIS		**98.45**
Monte Verde 44		**35.63**
Bosques de Alcalá		**61.54**
Emmanuel		**59.06**
Reserva de Zamorano 414 Etapa 2		**17.99**
Chapinero Sur VIS		**2.33**

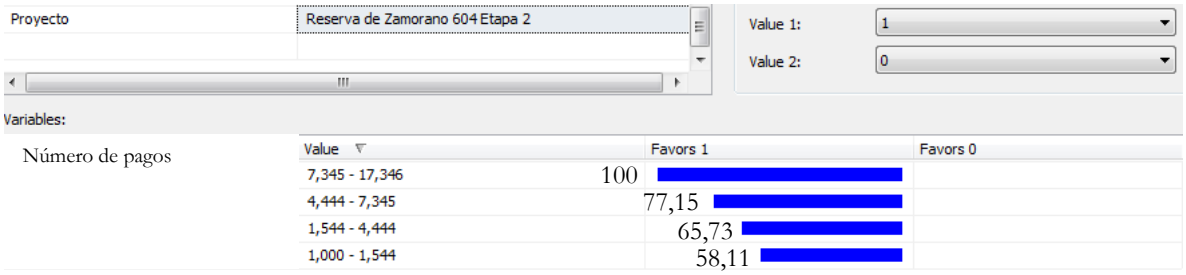
Fuente: Construcción propia

3.3.5. Regresión Logística

Con el fin de utilizar una técnica de regresión, fue necesario establecer si para el problema de investigación aplicaba la regresión lineal o la regresión logística, tal y como lo diferencian (Witten y Hall, 2011). La primera, se utiliza cuando se desea predecir un valor numérico, por ejemplo, el monto de las ventas; y la segunda si la variable de predicción es discreta, por ejemplo si una acción sube o no en una bolsa de valores. De esta manera se determinó que la variable de predicción (comprador o no comprador) es discreta.

La Figura 7 muestra un resultado para la regresión logística, donde la entrada de atributos se filtró con algunas características encontradas en la técnica de clustering, concretamente el clúster No. 3, que tenía casos con la mayor probabilidad de compra y que caracterizan un comprador como un individuo de edad media 42 años, que no es cabeza de hogar, y quien tiene en su historial de pagos un abono que llega a aproximadamente 21 millones de pesos. Con los datos anteriores se quiso conocer por cada proyecto, cuál sería el rango en el que oscila el número de cuotas que se esperan recibir de un comprador con esas características. El tamaño de la barra y su rótulo muestran la puntuación de importancia, donde el valor 1, representa una compra efectiva.

Figura 7: Ejecución de la técnica de regresión logística sobre el data warehouse, utilizando como datos de entrada, resultados parciales de la técnica de clustering



Fuente: Extraído de la herramienta Microsoft Analysis Servicies

3.3.6. Reglas de Asociación

Como lo afirman (Ma & Capri, 2014), una regla de asociación (RA) se define como una implicación de la forma Antecedente → Consecuente. Al hacer énfasis sobre el sistema de pagos del problema de investigación, se encontraron algunas reglas de asociación descritas como sigue:

Número de pagos = 6 - 11, Valor de venta < = 62.008.193 → Venta = Sí
Número de pagos = 6 - 11, Valor de venta < 62.008.193, Valor Cuota < 3.728.801 → Venta = Sí
Valor de venta < = 62.008.193, Valor Cuota < 3.728.801 → Venta = Sí
Pagos Totales > 45.104.500, Valor Cuota < 3.728.801 → Venta = Sí
Pagos Totales > 45.104.500, Valor de Venta > 62.008.193, Valor Cuota < 3.728.801 → Venta = Sí
Pagos Totales > 45.104.500, Valor Venta > 62.008.193 → Venta = Sí

Las reglas de asociación mostraron muchos resultados triviales, lo cual se debe a que la técnica tiene como objetivo el énfasis en el conocido caso de análisis de la cesta de mercado y en el problema de investigación se evidencia que la venta es entorno a un solo producto (vivienda).

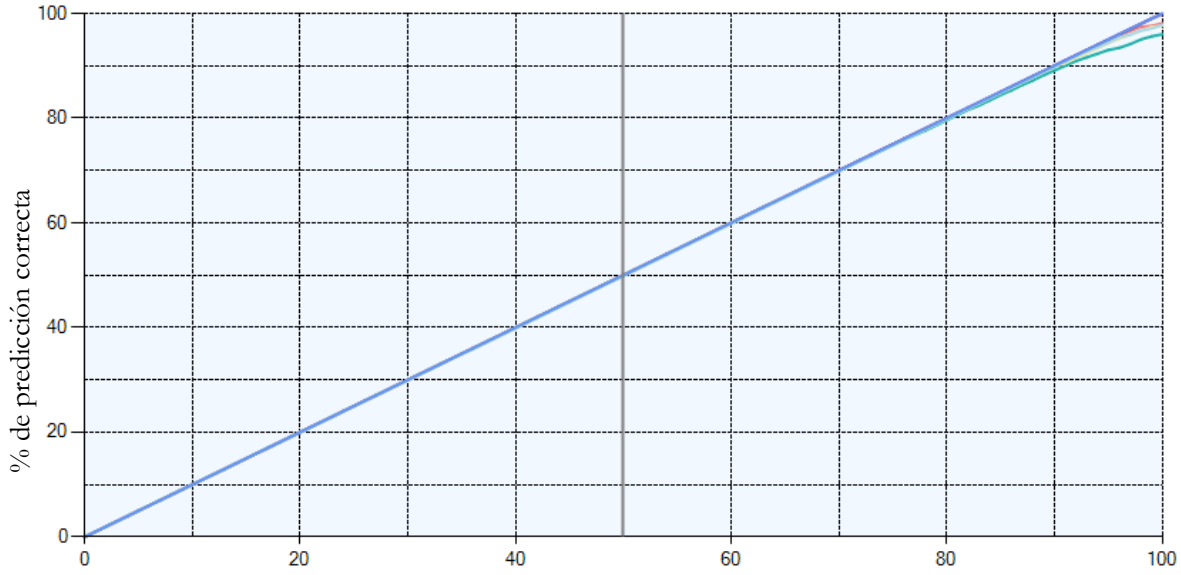
3.4. Evaluación de patrones y validación de los modelos

Antes de poder ratificar los patrones encontrados en la sección 3.3, se realizaron validaciones a dichos modelos. La técnica de reglas de asociación no fue tenida en cuenta debido a su modesto aporte al hallazgo de patrones y por ende, se validaron los modelos construidos sobre: árboles de decisión, clustering, redes neuronales, regresión logística y clasificación bayesiana.

3.4.1. Gráfico de elevación

Al comparar los puntajes de elevación para los diferentes modelos mostrado en la Figura 8, se puede observar la precisión que tienen usando el mismo atributo predecible (ventas). En esta primera prueba de validación todos los modelos presentan puntajes estrechamente similares.

Figura 8: Gráfico de elevación para 5 modelos de minería, aplicados sobre el data warehouse



Fuente: Extraído de la herramienta Microsoft Analysis Servicies

La línea diagonal representa un modelo ideal que a cualquier porcentaje de la población, predice correctamente la totalidad de los casos (máximo a esperar). La Tabla 2 relaciona el porcentaje de predicción correcta para cada modelo cuando se procesa un 95% de la población.

Tabla 2: Puntaje y porcentaje de predicción correcta de los modelos, al procesar un 95% de la población

Modelo	Puntaje	Población con predicción correcta
Árboles de decisión	1.0	94.90%
Clustering	0.99	93.00%
Clasificación Bayesiana	1.0	95.00%
Redes Neuronales	1.0	94.30%
Regresión Logística	1.0	94.40%

Fuente: Construcción propia

3.4.2. Matrices de Confusión

Los modelos presentan porcentajes de falsos positivos y falsos negativos (en conjunto) de: 1.9%, 4%, 0.1%, 2.2% y 2.4% para las técnicas de minería aplicadas en el orden de la Tabla 3, lo cual denota las falsas predicciones.

Tabla 3: Conteo de las matrices de confusión al ejecutar los modelos de minería con los datos de prueba

Modelo	Valor predicho	0 (valor real)	1 (valor real)
Árboles de Decisión	0	577	8
	1	11	404
Clustering	0	548	0
	1	40	412
Clasificación Bayesiana	0	588	1
	1	0	411
Regresión Logística	0	575	9
	1	13	403
Redes Neuronales	0	574	10
	1	14	402

Fuente: Construcción propia

3.4.3. Validación Cruzada

La validación cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba (Kusy & Kluska, 2017). Esta validación incluye pruebas de clasificación (éxito o fallo) y de probabilidad (elevación, puntuación logarítmica y error cuadrático). La Tabla 4, muestra un resumen de la validación cruzada en la prueba de clasificación, donde los datos de entrenamiento fueron divididos en 10 modelos temporales:

Tabla 4: Resultados de las pruebas de clasificación sobre los modelos de minería

Modelo	Éxito	Fallo
Árboles de decisión	98,51%	1,49%
Clustering	96,82%	3,18%
Clasificación Bayesiana	99,62%	0,38%
Redes Neuronales	97,23%	2,77%
Regresión Logística	97,05%	2,95%

Fuente: Construcción propia

Con el fin de medir la probabilidad de predicción de cada modelo se realizaron mediciones a través de la elevación, puntuación logarítmica y error cuadrático, sobre los datos de entrenamiento divididos en 10 modelos temporales (ver Tabla 5). La elevación corresponde a la relación entre la probabilidad de predicción real y la probabilidad marginal (cada modelo temporal). Esta medida normalmente muestra la mejora de la probabilidad del resultado de destino cuando se usa el modelo. A mayor valor, mejor es el modelo. Por su parte, la regla de puntuación logarítmica se utiliza para medir qué tan bien se realiza una asignación dada de probabilidades a los valores de una variable aleatoria en algunas instancias de dicha variable; cuanto menor sea el valor con la regla de puntuación logarítmica, mejor se realizará la asignación de probabilidades según la regla. Finalmente, el error cuadrático mide la cantidad de error que hay entre dos conjuntos de datos, comparando el valor predicho con el valor conocido, a menor error, mejor es el modelo.

Tabla 5: Pruebas de probabilidad sobre los modelos de minería de datos

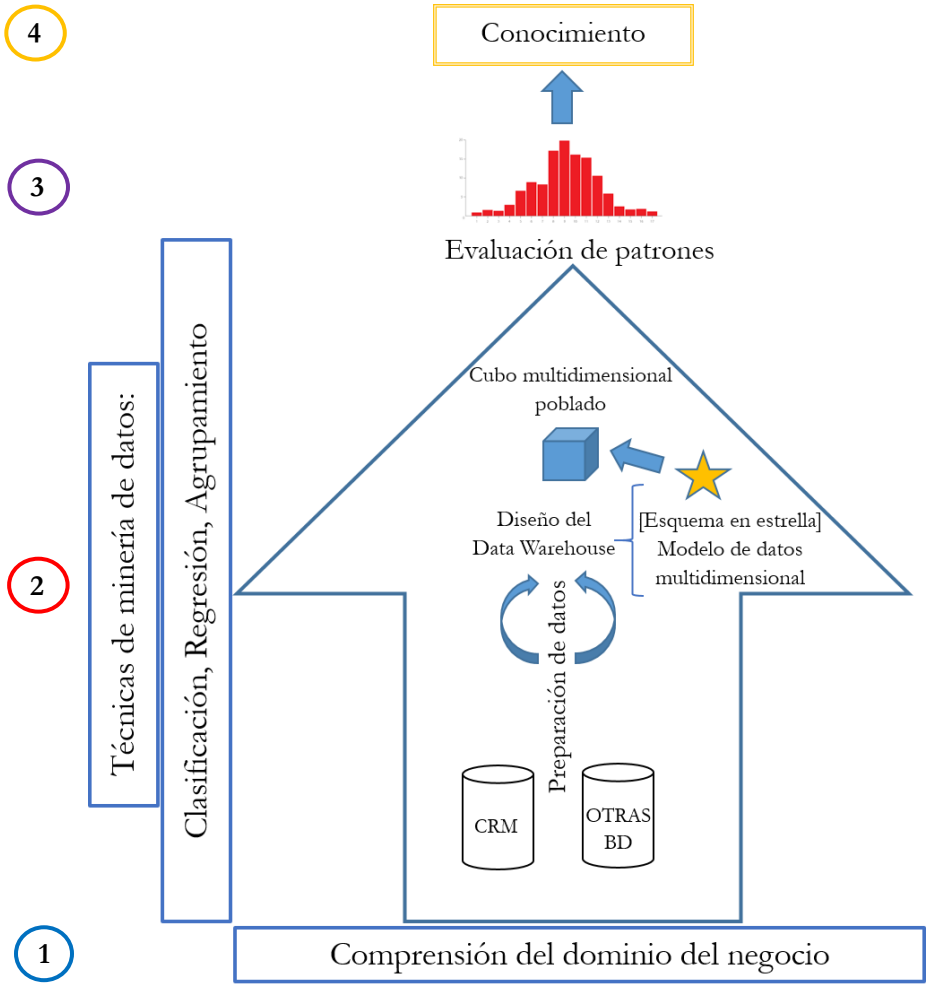
Modelo	Elevación (promedio)	Puntuación logarítmica (promedio)	Error cuadrático (promedio)
Árboles de decisión	0,62	-0,05	0,05
Clustering	0,59	-0.09	0,08
Clasificación Bayesiana	0,67	-0,01	0,05
Redes Neuronales	0,60	-0,07	0,06
Regresión Logística	0,57	-0,10	0,06

Fuente: Construcción Propia

3.5. Generalización: Presentación del modelo de caracterización del perfil del cliente

La Figura 9 ilustra el modelo propuesto para la caracterización del perfil del cliente que se vincula a los procesos de comercialización de viviendas en las constructoras utilizando minería de datos. En él se describe la metodología a seguir y la descripción de las especificidades a tener en cuenta.

Figura 9: Modelo de caracterización del perfil del cliente adaptado al sector de la construcción



Fuente: Construcción propia

A continuación se detallan cada una de las cuatro etapas propuestas en el modelo:

3.5.1. Comprensión del dominio del negocio

La comprensión del dominio del negocio debe incluir una contextualización exigente, acompañada de un trabajo de campo donde se desarrollen entrevistas con personal del área comercial, que aborden el nivel directivo y operativo, así como también conocer a una pequeña muestra de los clientes. Resulta de gran importancia la revisión de los indicadores de ventas de los que se tenga registro, así como también informes, históricos y reportes de otras áreas que interactúen con el proceso comercial y conocer los objetivos estratégicos propuestos por la dirección.

Los procesos de comercialización de vivienda también pueden incluir flujos de información con entidades externas, como bancos, cajas de compensación, notarías y registro. Para completar la comprensión del dominio del negocio es necesario identificar las interacciones que se dan con las entidades externas y abstraer las relaciones de datos con el proceso comercial.

3.5.2. Aplicar técnicas de minería de datos

En el modelo propuesto, esta fase consiste en ejecutar algoritmos sobre el conjunto de datos transformados en búsqueda de patrones. Para ello, se inicia con la preparación de datos que incluye la limpieza, la cual remueve ruido y

datos inconsistentes, esto se logra eliminando datos redundantes, corrigiéndolos o completándolos. Las fuentes de datos pueden corresponder a bases de datos genéricas o también bases de datos provenientes de software del tipo CRM. En el primer caso se requiere extraer la información del proceso comercial, descartando la de otros procesos de la empresa constructora; para el caso de los CRM's, se puede contar de manera inmediata con la información comercial de relacionamiento con el cliente. El resultado de la preparación de datos se ve reflejado en el diseño del data warehouse que se sugiere construir sobre un modelo de datos multidimensional basado en el esquema en estrella.

Una vez se accede a la estructura de la base de datos, se analizan cuáles entidades pueden contribuir a la caracterización del perfil del cliente, es decir, se acotan las entidades y relaciones al problema de investigación. Otros criterios para delimitar, tienen que ver con el análisis del volumen de datos y la cronología asociada a los datos por cada entidad.

Antes de poder aplicar las técnicas de minería de datos se debe poblar el data warehouse según el diseño (estructura) y es allí donde el cubo multidimensional poblado o con datos, se convierte en el insumo para la minería de datos. Las técnicas que se proponen son: clasificación, regresión y agrupamiento (clustering) bajo el paradigma de aprendizaje supervisado.

3.5.3. Evaluación de patrones

En esta etapa se realiza una minuciosa revisión de los patrones encontrados, bajo el principio que no todo patrón corresponde a un resultado válido y por ello los resultados se validan contra los objetivos del problema de investigación. Es necesario también realizar validaciones técnicas sobre los modelos empleando gráficos de elevación, matrices de confusión y la validación cruzada.

3.5.4. Conocimiento encontrado

La entrega del conocimiento a los grupos de interés debe realizarse a través de formas de representación fácilmente comprensibles a los usuarios, ya que generalmente desconocen el lenguaje técnico que se empleó en el proceso de minería de datos. En otras palabras, el conocimiento encontrado se debe exponer a nivel de informes gerenciales, en cuyo caso las áreas comerciales de las constructoras están familiarizadas con esta metodología de presentación de resultados. Finalmente es importante recomendar la manera en que este diamante de información pueda ser usado estratégicamente en el área comercial de la empresa constructora para que sea una herramienta para el mejoramiento continuo.

Los resultados presentados en esta sección van de lo particular a lo general. Si bien es cierto, se muestran resultados para el caso de estudio, también se expone el modelo propuesto para la caracterización del perfil del cliente que se vincula al proceso de comercialización de viviendas que otras empresas constructoras pueden adoptar para obtener resultados similares en el contexto del negocio y que les permita conocer el cliente objetivo, común en el sector de la construcción. Lo anterior corrobora la tendencia y necesidad de los sectores productivos de contar con modelos de minería de datos particulares a sus requisitos.

4. CONCLUSIONES

La minería de datos ofrece un marco de trabajo para descubrir conocimiento, pasando de tener datos a obtener información valiosa que no es visible a simple vista (conocimiento). No obstante, los problemas de investigación en minería de datos difieren en los diversos sectores económicos, lo que ha ofrecido la oportunidad de generar modelos particulares a los gremios dependiendo de las necesidades de búsqueda.

La transformación de los datos y el diseño de un data warehouse puede realizarse a partir de software CRM o a partir de bases de datos generales. El diseño de un cubo multidimensional con estructura tipo estrella, donde las dimensiones y hechos correspondan a entidades relacionadas con el perfil del cliente ofrece un buen insumo para la minería de datos en búsqueda de la caracterización del cliente objetivo que se vincula al proceso de comercialización de viviendas.

Las técnicas de minería de datos orientadas a la clasificación, regresión y agrupamiento con el enfoque de aprendizaje supervisado son las recomendadas para minar los datos del cliente del sector de la construcción, sin importar la herramienta de software que se utilice. Se propone que estas técnicas sean validadas a través gráficos de elevación, matriz

de confusión y la validación cruzada, lo cual garantiza la confiabilidad de las predicciones.

Una parte importante del análisis de resultados tiene que ver con que existe diversidad en los mismos, ofrecidos por las técnicas de minería de datos y sus parametrizaciones, obteniendo resultados de mayor y menor relevancia. Pese a que en el caso de estudio existían modelos con mejores puntajes en las pruebas de validación, en la presentación al cliente (área comercial) se evidenció complacencia por los resultados obtenidos con la técnica de clustering, ya que ofrece una caracterización de los clientes, sin desconocer los otros resultados igualmente valiosos. Con la caracterización del perfil del cliente que se vincula a los procesos de comercialización de vivienda en las constructoras, se puede llegar más fácilmente a los puntos de equilibrio de venta de los proyectos, lo que da luz verde a la construcción o ejecución de obra y por ende ofrece mejoras a la rentabilidad y la imagen corporativa.

Un trabajo futuro puede enfocarse en técnicas de aprendizaje no supervisado que introducen aún más el concepto de machine learning e inteligencia artificial. Una mayor automatización del proceso de minería de datos resulta acorde con el avance de los sistemas de información y las organizaciones, donde se requieren soluciones rápidas y adaptativas para la toma de decisiones.

REFERENCIAS

Abdellah, A., Rachid, A., & Belaid, B. (2016). Efficiency comparison and evaluation between two ETL extraction tools. *Indonesian Journal of Electrical Engineering and Computer Science Vol. 3, No. 1*, 174-181.

Adekitan, A., Abolade, J., & Shobayo, O. (2019). Data mining approach for predicting the daily internet data traffic of a smart university. *Journal of Big Data*, 6-11.

Al-Twijri, M., & Noamanb, A. (2015). A New Data Mining Model Adopted for Higher Institutions. *International Conference on Communication, Management and Information Technology. Procedia Computer Science 65*, 836 – 844.

Azzalini, A., & Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction.* Oxford: University Press.

Bahari, F., & Elayidom, S. (2015). An efficient CRM-Data Mining framework for the prediction of customer behavior. *Procedia Computer Science 46*, 725 – 731.

Cámara Colombiana de la Construcción CAMACOL. (2018). *Informe de actividad económica 2018: Colombia Construcción en Cifras.* Bogotá.

Dariush, F., Hojjatollah, S., Hajigol, E., & Parirooy, N. (2016). Classification of Bank Customers by Data Mining: a Case Study of Mellat Bank branches in Shiraz. *International Journal of Management, Accounting and Economics Vol. 3, No. 8*, 2383-2126.

Dutta, I., Dutta, S., & Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems With Applications 90*, 374-393.

GAIA TELCOM Sucursal Colombia y Consorcio Moreno Tafurt S.A. (2014). Contrato comercial de prestación de servicios tecnológicos.

Gajzler, M. (2016). Usefulness of mining methods in knowledge source analysis in the construction industry. *Gruyter Open Archives of Civil Engineering Vol. LXII* , SSUE I.

Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques, Third Edition.* University of Illinois: at Urbana-Champaign.

Kaya, M., Keles, A., & Oral, E. (2014). Construction Crew Productivity Prediction By Using Data Mining Methods. *Procedia - Social and Behavioral Sciences*, 1249 – 1253.

Kittipong, S., Sukree, S., & Pongpun, A. (2019). Application of Kansei engineering and data mining in developing an ingenious product co-design system. *International Journal of Machine Learning and Computing, Vol. 9, Issue 1*, 67-74.

Kusy, M., & Kluska, J. (2017). Assessment of prediction ability for reduced probabilistic neural. *Soft Computing 21*, 199-

212.

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons 60*, 293—303.

Listanti, M. (2014). Proactive Supply Chain Performance Management with Predictive Analytics. *The Scientific World Journal Volume 2014*, Article ID 528917.

Ma, X., & Capri, H. (2014). *Data mining: principles, applications and emerging challenges.* New York: Nova Science Publishers, Inc.

Matignon, R. (2007). *Data Mining Using SAS® Enterprise Miner.* San Francisco: Amgen, Inc.

Microsoft, Corporation. (12 de Abril de 2018). *About SQL Server Analysis Services*. Obtenido de https://docs.microsoft.com/en-us/sql/analysis-services/analysis-services?view=sql-server-2017

Naika, A., & Samantb, L. (2016). Correlation review of classification algorithm using data mining tool: Weka, Rapidminer, Tanagra, Orange and Knime. *International conference on computational modeling and security (CMS). Procedia Computer Science 85*, 662-668.

Patel, M. H., Abbasi, M. A., Saeed, M., & Alam, S. J. (2018). A scheme to analyze agent-based social simulations using exploratory data mining techniques. *Complex Adaptive Systems Modeling 6*, 1-17.

Rygielski, C., & Jyun-Cheng, W. (2002). Data mining techniques for customer relationship management. *Technology in Society 24*, 483–502.

Steffen, H., Hajo, W., Dorothea, S., & Steffen, I. (2019). DMME Data mining methodology for engineering applications –a holistic extension to the CRISP-DM model. *Procedia CIRP 79. 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering*, 403-408.

Vasoya, A., & Koli, N. (2016). Mining of association rules on large database using distributed and parallel computing. *7th International Conference on Communication, Computing and Virtualization. Procedia Computer Science 79*, 221-230.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques. Third Edition.* San Francisco: Morgan Kaufmann.

Yong, L., Jiecai, Z., & Xueqing, L. (2018). An Incremental Association Rule Algorithm Based on MapReduce. *Journal of Physics Conf. Series 1069*, 012102.