ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÍSTER EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA: MODELO DE ANÁLISIS PREDICTIVO PARA EL MEJORAMIENTO PRESUPUESTARIO DE LA PLANIFICACIÓN LOGÍSTICA DEL CONSEJO NACIONAL ELECTORAL DEL ECUADOR, BASADO EN EL USO DE TÉCNICAS DE MINERIA DE DATOS

AUTOR: ING. ORTUÑO ULCO, MARÍA TEODORA

DIRECTOR: MGTR. NINAHUALPA QUIÑA, GEOVANNI

SANGOLQUÍ

2019

CERTIFICADO DEL DIRECTOR

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"MODELO DE ANÁLISIS PREDICTIVO PARA EL MEJORAMIENTO PRESUPUESTARIO DE LA PLANIFICACIÓN LOGÍSTICA DEL CONSEJO NACIONAL ELECTORAL DEL ECUADOR, BASADO EN EL USO DE TÉCNICAS DE MINERIA DE DATOS"** fue realizado por la Ing. **MARÍA TEODORA ORTUÑO ULCO** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 12 de diciembre de 2018

GEOVANNI NINAHUALPA QUIÑA

C.C 1709036261

AUTORÍA DE RESPONSABILIDAD

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORÍA DE RESPONSABILIDAD

Yo, **MARÍA TEODORA ORTUÑO ULCO**, con cédula de ciudadanía N°1715319131, declaro que el contenido, ideas y criterios del trabajo de titulación: **"MODELO DE ANÁLISIS PREDICTIVO PARA EL MEJORAMIENTO PRESUPUESTARIO DE LA PLANIFICACIÓN LOGÍSTICA DEL CONSEJO NACIONAL ELECTORAL DEL ECUADOR, BASADO EN EL USO DE TÉCNICAS DE MINERIA DE DATOS"** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 12 de diciembre de 2018

MARÍA TEODORA ORTUÑO ULCO

C.C.: 1715319131

AUTORIZACIÓN

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN

Yo, **MARÍA TEODORA ORTUÑO ULCO** autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación **"MODELO DE ANÁLISIS PREDICTIVO PARA EL MEJORAMIENTO PRESUPUESTARIO DE LA PLANIFICACIÓN LOGÍSTICA DEL CONSEJO NACIONAL ELECTORAL DEL ECUADOR, BASADO EN EL USO DE TÉCNICAS DE MINERIA DE DATOS"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 12 de diciembre de 2018

MARÍA TEODORA ORTUÑO ULCO

C.C.: 1715319131

DEDICATORIA

El presente trabajo de investigación es dedicado a Dios, por darme la oportunidad de tener una vida llena de aprendizajes, experiencias y regalarme una familia maravillosa, por fortalecer mi corazón y guiarme a lo largo de mi carrera. A mis padres quienes son mi motor y mi mayor inspiración, que a través de su amor, paciencia, buenos valores, ayudan a trazar mi camino. A mi esposo e hijos por sus palabras y confianza, por su amor, cariño y brindarme el tiempo necesario para realizarme profesionalmente, a mis amigos, compañeros y a todas aquellas personas que de una u otra manera han contribuido para el logro de mis objetivos. A mi tutor de tesis, por su valioso asesoramiento en la realización de esta investigación.

AGRADECIMIENTO

El presente trabajo está dedicado a todas aquellas personas que, de alguna forma hicieron posible esta investigación y son parte de su culminación. Mi sincero agradecimiento está dirigido a mi familia por brindarme siempre su apoyo y cariño, pero, principalmente mi agradecimiento a mi querido esposo por ser el apoyo incondicional en mi vida, que, con su amor y respaldo, me ayuda alcanzar mis objetivos. A mis docentes a quienes les debo gran parte de mis conocimientos, gracias a su paciencia y enseñanza y finalmente un eterno agradecimiento a esta prestigiosa universidad que me ha permitido concluir con una etapa más de mi vida para un futuro competitivo.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El Consejo Nacional Electoral (CNE) es el máximo organismo de sufragio del Ecuador, sus funciones son organizar, controlar las elecciones, sancionar a partidos y candidatos que infrinjan las normas electorales; además inscriben y fiscalizan a los partidos y movimientos políticos. Actualmente, esta institución específicamente la Dirección Nacional de Logística presenta una inadecuada proyección del presupuesto del material electoral, realizando este proceso mediante la experiencia de la dirección mencionada sin ninguna técnica que sustente dicho valor, el proceso lo realizan mediante cálculos y consolidaciones de datos en hojas Excel, por lo que el trabajo toma mucho tiempo, tanto de procesamiento y tratamiento de datos. El presente proyecto tiene como propósito mejorar la planificación del proceso electoral, a través del diseño de un modelo analítico para la predicción del presupuesto de papeletas electorales definido de acuerdo a un estudio de técnicas, herramientas y algoritmos analíticos más utilizados en el mercado. Se utilizaron los tipos de investigación exploratoria y descriptiva para recopilar los datos del negocio; estas técnicas fueron aplicadas siguiendo los lineamientos de la metodología de desarrollo CRISP –DM específica para minería de datos. Los resultados muestran que se obtiene una precisión del 89,58 %, a diferencia del 25% obtenido antes de aplicar la solución propuesta. En consecuencia, podemos avizorar la efectividad del modelo, por lo que su aplicación permitirá un análisis eficiente en la toma de decisiones, y de esta forma mejorar la proyección presupuestaria y la optimización del recurso económico para las demás direcciones de la institución.

PALABRAS CLAVE:

- **INDICADORES PRESUPUESTARIOS**
- **MODELOS PREDICTIVOS**
- **PATRONES DE COMPORTAMIENTO**

ABSTRACT

The National Electoral Council (CNE) is Ecuador's highest suffrage body, its functions are to organize, control elections, sanction parties and candidates that violate the electoral norms; they also inscribe and supervise political parties and movements. Currently, this institution specifically the National Logistics Office presents an inadequate projection of the electoral material budget, carrying out this process through the experience of the aforementioned management without any technique that supports this value, the process is done through calculations and consolidations of data in sheets Excel, so the work takes a lot of time, both processing and processing data. The purpose of this project is to improve the planning of the electoral process, through the design of an analytical model for the prediction of the electoral ballot budget defined according to a study of techniques, tools and analytical algorithms most used in the market. The types of exploratory and descriptive research were used to collect the business data; these techniques were applied following the guidelines of the CRISP development methodology -DM specific for data mining. The results show that an accuracy of 89.58% is obtained, unlike the 25% obtained before applying the proposed solution. Consequently, we can envisage the effectiveness of the model, so that its application will allow an efficient analysis in decision-making, and in this way improve the budgetary projection and the optimization of the economic resource for the other directions of the institution.

KEYWORDS:

- **BUDGET INDICATORS**
- **PREDICTIVE MODELS**
- **BEHAVIOR PATTERNS**

CAPITULO I

INTRODUCCIÓN

1.1 Antecedentes

El artículo 217 de la Constitución de la República del Ecuador, en concordancia con lo establecido en el artículo 18 de la ley Orgánica Electoral y de Organizaciones Políticas de la República del Ecuador, Código de la Democracia, establece que: "la Función Electoral garantizará el ejercicio de los derechos políticos que se expresan a través del sufragio, así como los referentes a la organización política de la ciudadanía".

El artículo 7 del artículo 219 de la Constitución de la República del Ecuador, en concordancia con el numeral 10 del artículo 25 de la Ley Orgánica Electoral y de Organizaciones Políticas de la República del Ecuador, Código de la Democracia, establecen como una de las funciones del Consejo Nacional Electoral la de "Determinar su organización, formular y ejecutar su presupuesto".

El Estatuto Orgánico de Gestión Organizacional por Procesos del Consejo Nacional Electoral establece que la misión de la Gestión Nacional de logística, es: "Planificar, organizar y ejecutar los medios y métodos necesarios para proveer, desarrollar y hacer seguimiento de manera oportuna del material electoral requerido por el Consejo Nacional Electoral cumpliendo los estándares y normas de calidad para el desarrollo de los procesos electorales y otros establecidos en la ley"; y, su responsable es el/la Directora(a) Nacional de Logística.

Para ejecutar el mandato que se le atribuyó en el Estatuto Orgánico de Gestión por Procesos del Consejo Nacional Electoral como administrador único del material electoral requerido por el mismo, la Dirección Nacional de Logística realiza la proyección presupuestaria del material electoral en cada proceso.

Por la dimensión de los procesos que se maneja en la Dirección Nacional de Logística, se dispone de una gran cantidad de datos dispersos y que al momento de tomar una decisión no se encuentran disponibles ya que no poseen una herramienta que consolide estos datos y reporte información precisa, actualizada y de utilidad que apoye este tipo de actividad, causando así mayor dificultad al tomar una decisión y llegar a un consenso entre las personas encargadas de hacerlo ya que cada uno tiene su propia perspectiva al analizar los datos.

Como un aporte al sector público en especial al CNE, el desarrollo de este trabajo está orientado a crear un modelo de análisis predictivo que permita el análisis de la información que involucra la toma decisiones en el sector público dirigido a la proyección presupuestaria considerando que al momento de tomar una decisión se encuentran involucrados factores como económicos, la complejidad del modelo organizacional y el personal que labora en la misma. En este escenario es necesario contar con un análisis eficiente de la información, que sea actualizada, precisa y que sirva de apoyo para el adecuado direccionamiento estratégico.

1.2 Justificación e Importancia

La Dirección Nacional de Logístico del CNE, cumple con funciones y ejecuta proyectos dentro de los cuales se encuentra como una actividad la proyección del presupuesto del material electoral, los mismos que no han logrado tener los resultados y beneficios esperados, como lograr que esta proyección sea efectiva, es decir que el presupuesto se ajuste lo más posible a lo que realmente se ejecuta.

Por tal razón, deben orientarse los recursos tecnológicos de manera que ayuden a tomar decisiones estratégicas y oportunas ante la ausencia de un modelo que facilite la toma de decisiones respecto a la proyección del presupuesto del material electoral. El conocimiento del comportamiento histórico de las proyecciones de los presupuestos podría ser de suma importancia en la elaboración y desagregación del presupuesto para los procesos futuros.

Es importante considerar la construcción y análisis de un modelo en instituciones públicas financieras en donde la implementación de análisis sustentados en minería de datos está iniciando. La presente idea puede ser considerada como un prototipo modelo de análisis predictivo, que pueden servir de base para el análisis en otras entidades del sector financiero contribuyendo al marco referencial en investigaciones de proyectos de minería de datos, así como también generará un conocimiento nuevo que servirá de apoyo en la toma de decisiones de las autoridades competentes de la institución.

1.3 Planteamiento del problema

La Dirección Nacional de Logística del Consejo Nacional Electoral, es la encargada de realizar la proyección del presupuesto del material para cumplir con los procesos electorales que el Estado tiene planificados en base a la Constitución de la Republica, este proceso genera gran volumen de información. Sin embargo, esta información no es utilizada como referente en la elaboración del presupuesto, este proceso solamente se lo realiza basado en el conocimiento del personal de logística cuyo acierto promedio es del 25%, ya que los especialistas encargados de esta tarea se basan fundamentalmente en datos que vienen recopilando de forma manual en hojas de Excel y en el análisis de los porcentajes de ejecución de las partidas del presupuesto del proceso anterior para realizar sus proyecciones, por lo que los cálculos actuales toman mucho tiempo de procesamiento, tratamiento y posteriormente esto conlleva a un análisis no adecuado de los datos y además un uso

deficiente de herramientas tecnológicas, generando por lo tanto un presupuesto sobredimensionado, que se refleja en una mala distribución del presupuesto electoral muy lejos a la realidad, provocando que otras direcciones de esta entidad, tengan que ajustar sus presupuestos y retrasar el cumplimiento de los objetivos fijados por cada una de las direcciones. Por otro lado, el Consejo Nacional Electoral con el valor del presupuesto devengado analiza la ejecución presupuestaria real, generando una devolución de recursos a las arcas fiscales del Ministerio de Economía y Finanzas. Para posteriores asignaciones presupuestarias, este ministerio realiza una evaluación de la planificación y ejecución presupuestaria de la Institución en relación a los valores históricos. Con esta evaluación el Ministerio de Economía Finanzas tiene la potestad de asignar o no el valor o recurso presupuestario solicitado por la Institución. El proceso descrito se ejemplifica en la Figura 1.

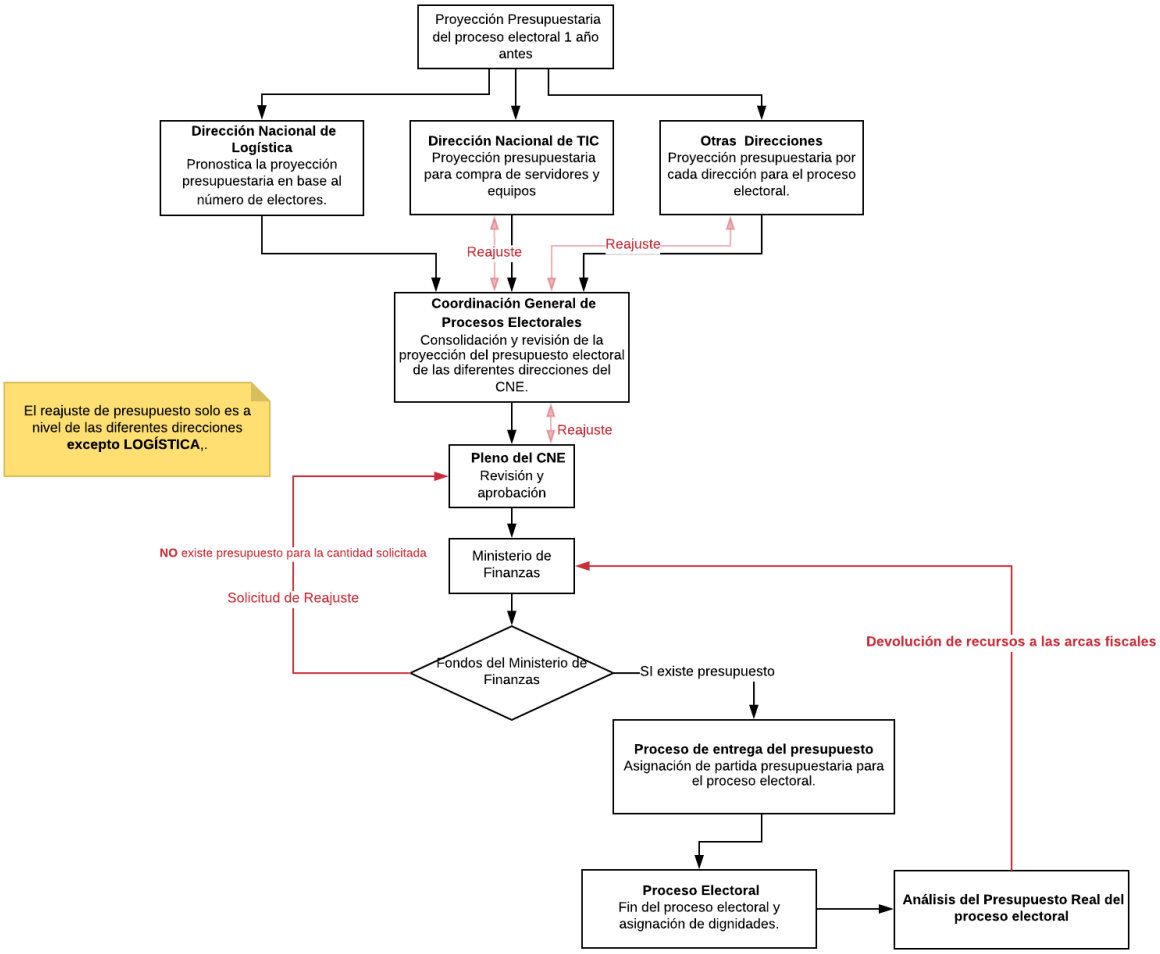

Figura 1. Proceso de planificación del presupuesto electoral

Por consiguiente, la Dirección Nacional de Logística busca ser eficiente en la elaboración del presupuesto del material electoral y para ello se requiere mejorar los procesos que actualmente disponen.

En el presente trabajo de investigación se pretende realizar un modelo predictivo para la proyección del presupuesto del material electoral, lo que permitirá una eficiente planificación presupuestaria y gestión efectiva de la Dirección Nacional de Logística.

En base a la problemática analizada, se plantea la siguiente pregunta de investigación:

¿Es posible mejorar la proyección de indicadores presupuestarios del material electoral en el Consejo Nacional Electoral del Ecuador, mediante la construcción de un modelo de predictivo basado en la obtención de patrones de comportamiento?

1.4 Objetivos

1.4.1 Objetivo general

Construir un modelo de predicción mediante la obtención de patrones de comportamiento para la proyección de indicadores presupuestarios del material electoral en el Consejo Nacional Electoral del Ecuador - CNE.

1.4.2 Objetivos específicos

- Realizar una revisión de literatura para determinar las herramientas, técnicas predictivas y modelos analíticos más utilizadas de minería de datos en el sector financiero, los mismos que aportarán en el planteamiento del modelo de predicción.
- Construir un modelo analítico-predictivo para la proyección del presupuesto, a través del uso de técnicas de minería de datos.
- Validar el modelo predictivo del presupuesto electoral, mediante el uso de técnicas de validación implementadas en minería de datos, para comprobar el nivel de confianza.

1.5 Hipótesis

Las técnicas de patrones de comportamiento permiten mejorar las predicciones en los indicadores presupuestarios del material electoral que maneja la Dirección Nacional de Logística.

Para la comprobación de la hipótesis planteada se considera el método deductivo usando metodologías de investigación como la entrevista, o la comprobación estadística, para posteriormente evaluar los resultados de la predicción a través de fórmulas de precisión del pronóstico de ventas utilizando cualquier error porcentual.

Para comprobar el modelo, se dividirá el conjunto de datos en dos subconjuntos, uno para el entrenamiento del modelo y el otro para la validación del mismo.

1.6 Alcance

Para el presente estudio se analizarán los datos de los presupuestos del material electoral para papeletas electorales de los procesos del 2002 y 2018 que mantiene el Consejo Nacional Electoral.

Con la consideración anterior, el alcance de la presente investigación es realizar un estudio comparativo, entre las técnicas, herramientas y algoritmos analíticos para encontrar el mejor modelo en la toma de decisiones de indicadores presupuestarios. Se considera también, los análisis predictivos basados en el cálculo de tendencias y posibilidades futuras, en la predicción de posibles resultados y en la formulación de recomendaciones. La construcción de un modelo predictivo en la Dirección Nacional de Logística del Consejo Nacional Electoral permitirá alertar con anticipación los resultados de los indicadores presupuestarios, de esta manera se contará con un análisis adecuado de los datos al momento de la elaboración de la proyección presupuestaria del material electoral.

1.7 Metodología

Para la definición de la metodología que se utilizará en el presente estudio, se realizó un mapeo de literatura en la base digital de Scopus con el objeto de determinar cuáles son las más usadas en proyectos de minería de datos. Adicionalmente se utilizó la técnica de Focus Group para conocer

la opinión de expertos en esta área. En base a este análisis se determinó que las metodologías KDD, SEMMA y CRISP-DM son las más utilizadas en el desarrollo de investigaciones basadas en minería de datos.

Como se puede visualizar en la Figura 2, los estudios que utilizan estas metodologías crecen con el transcurso del tiempo. Por tal motivo, realizaremos una revisión teórica de las mismas en el presente trabajo.

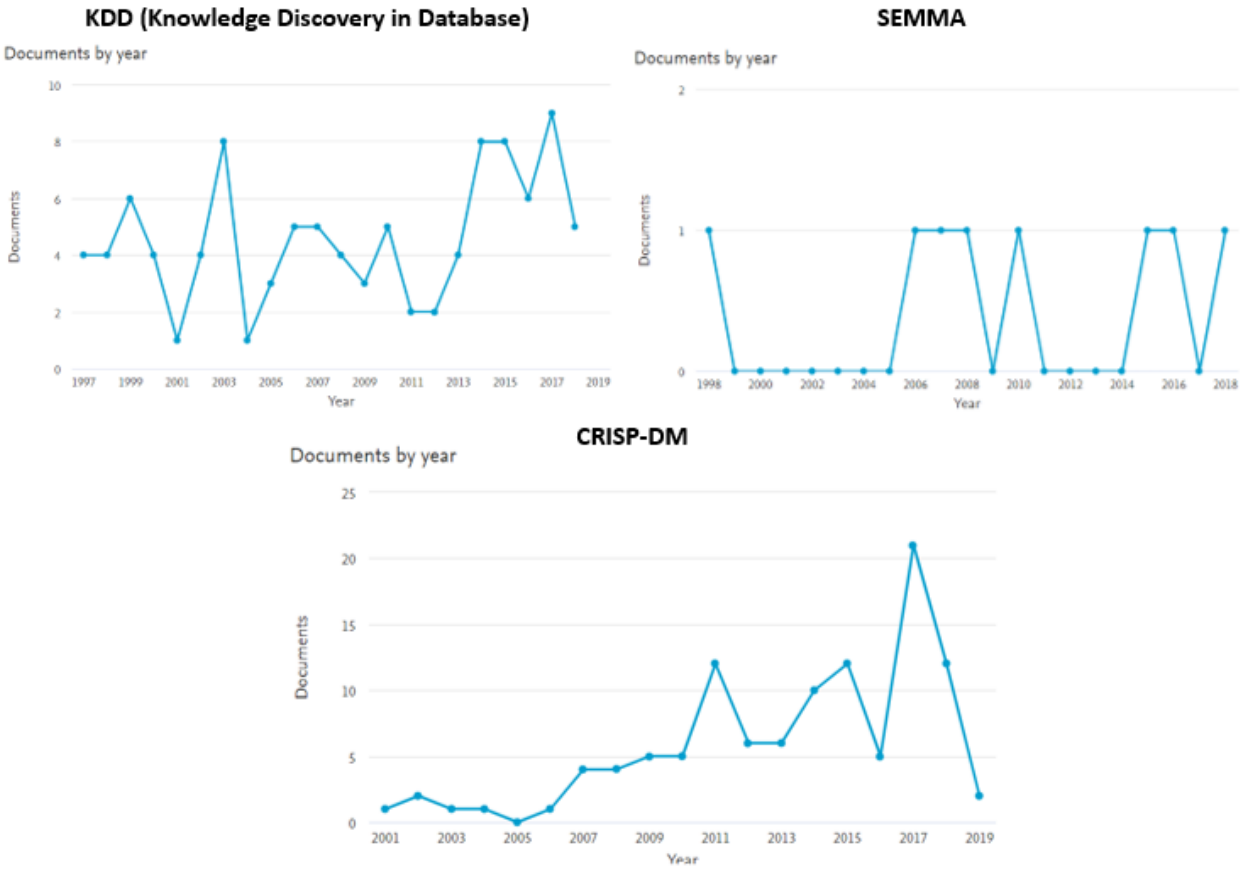

Figura 2. Tendencia de investigaciones que usan KDD, SEMMA y CRISP-DM en base digital Scopus
Fuente: *(Scopus, 2018)*

Revisión de las Metodologías KDD, SEMMA Y CRISP-DM

A principios del año 1996, KDD (Knowledge Discovery in Databases) constituyó el primer modelo aceptado en la comunidad científica que constituyo las etapas principales de un proyecto de explotación de información. Formalmente el modelo establece que la minería de datos es la etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos. Sin embargo, actualmente, en la comunidad científica y en la literatura, el término KDD y minería de datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento (Moine et al., 2011).

A inicios del año 2000, con el gran crecimiento que surgió en el área de la minería de datos, surgen tres nuevas metodologías que trazan un enfoque para llevar a cabo el proceso: SEMMA, Catalyst (conocida como P3TQ) y CRISP-DM.

Como se puede observar en la Figura 3, CRISP-DM se ha convertido en la metodología más utilizada, según un estudio publicado en el año 2014 por la comunidad KDnuggets (Data Mining Community's Top Resource).

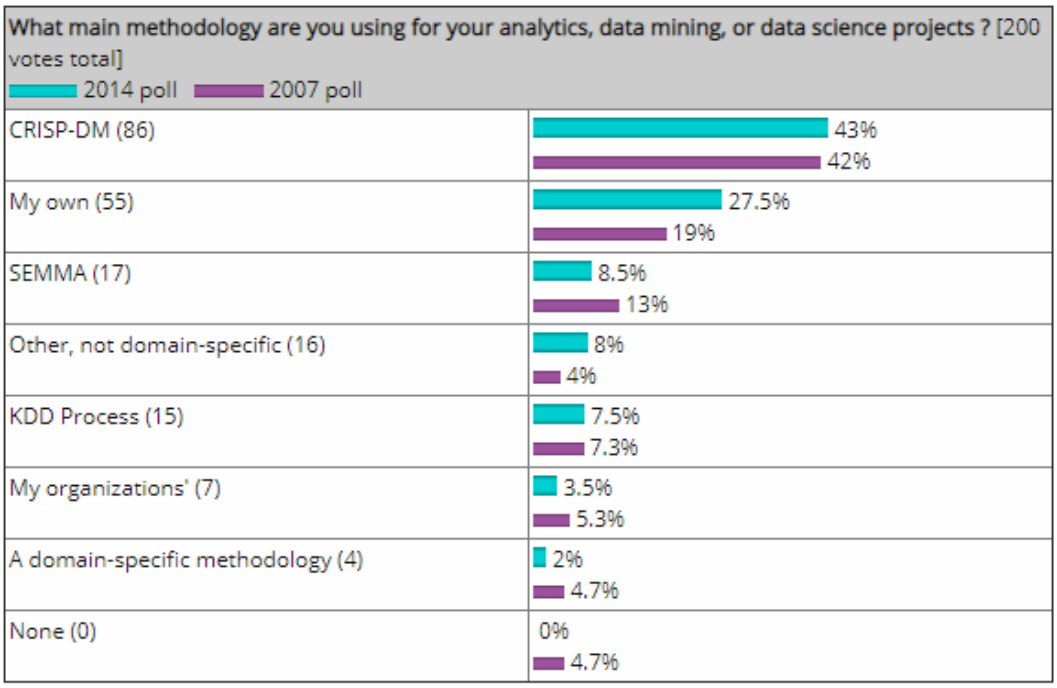

Figura 3. Encuesta realizada por KDnuggets en el año 2007 y 2014
Fuente: *(KDnuggets, 2014)*

Algunas metodologías profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otros proveen sólo una guía general del trabajo a realizar en cada fase (como el proceso KDD o SEMMA).

Adicionalmente, un estudio publicado en septiembre de 2016 referente a un análisis de las metodologías de minería de datos más utilizadas en el mercado, plantean a CRISP-DM, KDD y SEMMA dentro de su investigación como las de mayor popularidad. El objetivo del estudio fue realizar un comparativo de las interrelaciones entre las fases de las metodologías, como se observa en la Figura 4, en donde se puede visualizar que la Metodología CRISP-DM abarca todas las fases de las demás metodologías, además que contiene fases adicionales como: comprensión del negocio e implantación.

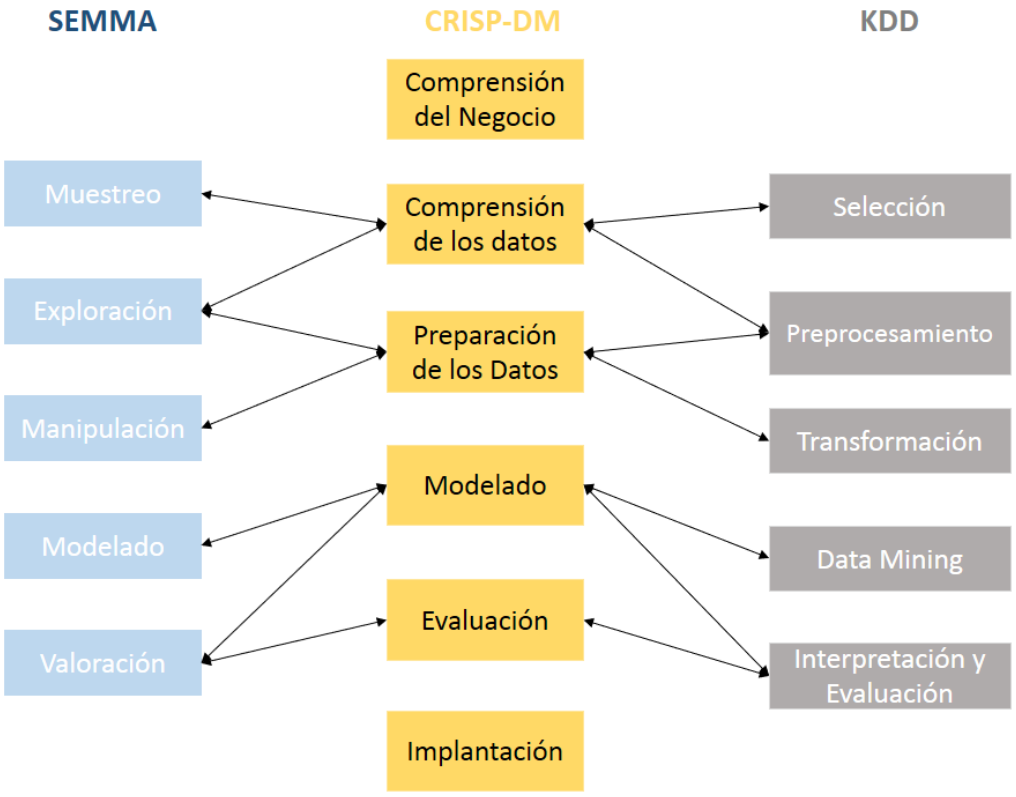

Figura 4. Comparativa de las interrelaciones entre las fases de las metodologías SEMMA, KDD y CRISP-DM
Fuente: (Rodriguez, Álvarez, Mesa, & González, , 2003)

A continuación, se describe una breve definición de cada metodología:

KDD, se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. Definitivamente no es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar analogías. Es un proceso que extrae información de calidad que puede

usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos (WebMining Consultores, 2011).

SEMMA, creada por SAS (Statical Analysis System), se define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos.

El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración).

La metodología SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS (Moine et al., 2011).

CRISP–DM (Cross Industry Standard Process for Data Mining), creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. Cada fase es descompuesta en varias tareas generales de segundo nivel (Moine et al., 2011).

Selección de la metodología

En base a la revisión de literatura antes descrita, se determinaron las metodologías de Data Mining más utilizadas y las características más importantes que deben contener, para determinar de manera objetiva aquella metodología que mejor se adapte para el desarrollo del presente trabajo. Para esta valoración y la determinación de los criterios de análisis, se ha utilizado la técnica de

focus group conformado por 4 profesionales con alta experiencia realizando proyectos de Inteligencia de Negocios, Big data y Analítica avanzada. La evaluación de las metodologías de minería de datos hecha por los participantes se encuentra detallado en la Tabla 1.

Tabla 1

Comparación de metodologías de minería de datos

	CARACTERÍSTICAS	**Metodología de Data Mining**		
		CRISP-DM	KDD	SEMMA
1	Robusta	✓	X	X
2	Baja Complejidad	✓	✓	✓
3	No tiene relación con herramientas comerciales	✓	✓	X
4	Menor curva de Aprendizaje	✓	✓	✓
5	Mayor número de fases	✓	X	X
6	Casos de éxito	✓	✓	X
7	Preferencia de los Data Science	✓	X	X

Los resultados generales de la valoración realizada por el focus group se visualiza en la Figura 5, en donde se puede observar que la metodología mejor evaluada es CRISP-DM, por tal motivo la misma será utilizada en el presente trabajo de investigación.

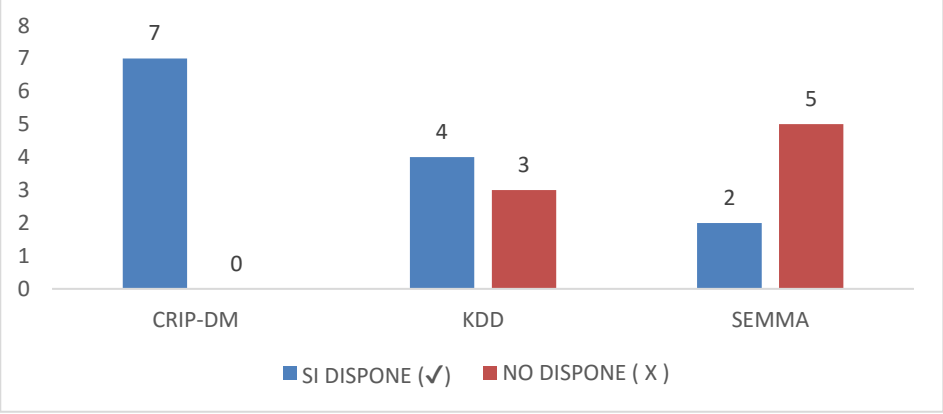

Figura 5. Evaluación de las características de las metodologías SEMMA, KDD y CRISP-DM

Descripción de la metodología seleccionada

Como se ha indicado en el apartado anterior, la metodología a aplicarse en el presente estudio será **CRISP-DM,** la misma que fue constituida como resultado de un proyecto en colaboración entre representantes de diferentes sectores industriales centrándose en su experiencia práctica implementando proyectos de Data Mining. El proyecto nació en 1996 liderado por un consorcio de tres empresas, Daimler-Benz (posteriormente DaimlerChrysler), ISL (posteriormente SPSS) y NCR, y financiado por la Comisión Europea. Tras sucesivas reuniones con otros profesionales del mismo campo, sesiones de trabajo y diversas implantaciones piloto, el proyecto concluyó en 1999 con un primer bosquejo de referencia, que posteriormente fue analizado hasta elaborar su primera versión completa (CRISP-DM 1.0) publicada en 2000.

El siguiente diagrama muestra las fases principales dentro del modelo de referencia CRISP-DM:

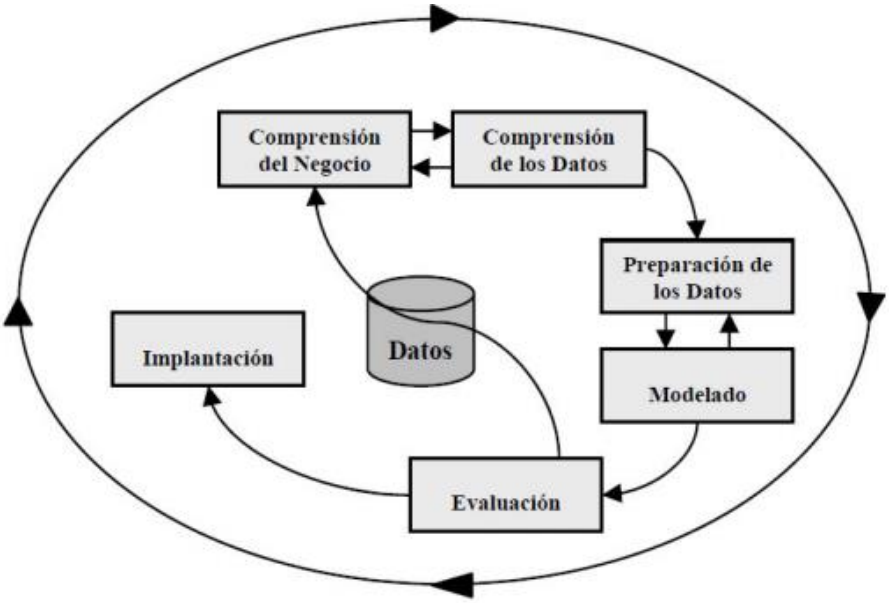

Figura 6. Modelo de proceso de la Metodología CRISP–DM
Fuente: *(Díaz, 2016)*

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

Comprensión del negocio (Objetivos y requerimientos desde una perspectiva no técnica), en esta fase se determinan las siguientes necesidades: Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito), Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio), Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito) y Generación del plan del proyecto (plan, herramientas, equipo y técnicas).

Comprensión de los datos (Familiarizarse con los datos teniendo presente los objetivos del negocio), principalmente en esta fase se realiza la recopilación inicial de datos, descripción de los datos, exploración y verificación de calidad de datos

Preparación de los datos (Obtener la vista minable o dataset), en esta fase se realiza las siguientes tareas: Selección de los datos, Limpieza de datos, Construcción de datos, Integración de datos y Formateo de datos.

Modelado (Aplicar las técnicas de minería de dato a los dataset), en esta fase se debe realizar la selección de la técnica de modelado, diseño de la evaluación, construcción del modelo y evaluación del modelo.

Evaluación (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio), en esta fase se realiza tareas como: evaluación de resultados, revisar el proceso y establecimiento de los siguientes pasos o acciones.

Despliegue (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización), en esta última fase se realiza la planificación de despliegue, planificación de la monitorización y del mantenimiento, generación de informe final y la revisión del proyecto.

Por otro lado, existen algunos proyectos de investigación publicados en SCOPUS y otros que han sido considerados casos de éxito como se visualiza en la Tabla 2, los mismos que comprueban que la metodología CRISP-DM de Data Mining está siendo utilizada por los investigadores y que su aplicabilidad sigue creciendo con el transcurso del tiempo, por tal motivo en el presente trabajo de investigación se aplicará la metodología CRISP-DM.

Tabla 2

Proyectos y caso de éxito aplicando la metodología CRISP-DM

Proyecto	Descripción
El desarrollo de un sistema de evaluación de desempeño usando el análisis del Árbol de Decisión y la Lógica Difusa	La investigación presenta el desarrollo de un sistema de evaluación del rendimiento que tiene como objetivo estudiar los recursos humanos específicos para el entorno educativo y pone de manifiesto el papel de la minería de datos para lograr un desarrollo mejorado de calidad. Los investigadores utilizaron metodologías CRISP-DM y Extreme Programming, centrándose en la generación de modelos para el algoritmo Decision Tree, combinado con Fuzzy Logic Controll[illegible] rendimiento de la facultad dentro de la universidad (Lamarca & Ambat, 2018).
Perfiles de pacientes con trombosis para determinar similitudes de comportamiento	Este caso muestra el proceso que se siguió para sacar perfiles de pacientes que tienen trombosis y que tienen pronósticos o comportamientos similares. La tarea de minería de datos que se utilizó en este caso fue la de segmentación y se utilizó la herramienta DB2 Intelligent Miner for Data para el desarrollo del proyecto de IBM utilizando la metodología CRISP-DM (Nishizaki Fernandes, 2017).
Análisis de datos para diagnóstico de síndrome metabólico	El objetivo de la investigación es proporcionar una revisión bibliográfica del estado actual de la técnica en el área del diagnóstico del síndrome metabólico mediante métodos de minería de datos. La

CONTINÚA

	estructuración de la revisión de la literatura está realizada mediante la metodología CRISP-DM, que generalmente se utiliza para organizar el proceso analítico (Pusztová, Babič, & Paralič, 2018).
Asociación de productos de la canasta de mercado para analizar el comportamiento de los clientes	Este caso se analiza el comportamiento de los clientes de una cadena de supermercados de acuerdo a sus compras. Se trata la tarea de asociación y se utilizó la herramienta de SAS, llamada Enterprise Miner en conjunto con la metodología CRISP-DM (Nishizaki Fernandes, 2017).
Caso de éxito: Banco BCI	Banco BCI formalizó y aceleró los procesos de desarrollo, evaluación y puesta en producción de modelos de riesgo de crédito gracias a la plataforma analítica de Smart Risk utilizando la metodología CRISP-DM (Solutions, 2017).

CAPITULO II

MARCO TEÓRICO

En este capítulo, se realizó una revisión de literatura de las principales herramientas, técnicas y modelos analíticos utilizados en el mercado para la proyección del presupuesto, con la finalidad de tener una idea de la tendencia del mercado y poder generar una propuesta óptima y verificable. Para la revisión de literatura se consideró lo siguiente:

- **Estudios de técnicas de minería de datos:** Para la revisión de técnicas de minería de datos se ha utilizado la información disponible en la web.
- **Estudios de algoritmos analíticos:** Para la revisión de algoritmos analíticos se utilizó KDnuggets sitio líder en Business Analytics, Big Data, Data Mining, Data Science y Machine Learning, empresa encargada de realizar investigaciones, encuestas para la verificación de preferencias en la elección de algoritmos.
- **Estudios realizados por empresas especializadas:** Revisión de estudios realizados por empresas especializadas en investigar las fortalezas y debilidades de los productos de fabricantes, soluciones y herramientas, así como las tendencias del mercado. Por lo anterior, se menciona el análisis realizado por las empresas: Gartner y Forrester para un análisis de herramientas analíticas; se considera estas empresas de investigación debido a que varios expertos en data mining lo recomiendan, uno de ellos es Gregory Piatetsky-Shapiro, Ph.D fundador de las conferencias Knowledge Discovery in Database (KDD) con más de 60 publicaciones, más de 10,000 citas, incluyendo 2 libros más vendidos y varias colecciones editadas sobre temas relacionados con la minería de datos y el descubrimiento de conocimiento (Kdnuggets, 2018).

2.1 Estudios de técnicas de minería de datos

La clasificación inicial de las técnicas de minería de datos distingue entre: **técnicas predictivas o aprendizaje supervisado**, en las que las variables pueden clasificarse en dependientes e independientes; **técnicas descriptivas o aprendizaje no supervisado**, en las que todas las variables tienen el mismo estatus; y técnicas auxiliares, en las que se realiza un análisis multidimensional de datos, visualizados en la Figura 7 (Santín Gonzalez & Peréz Lopéz, 2008).

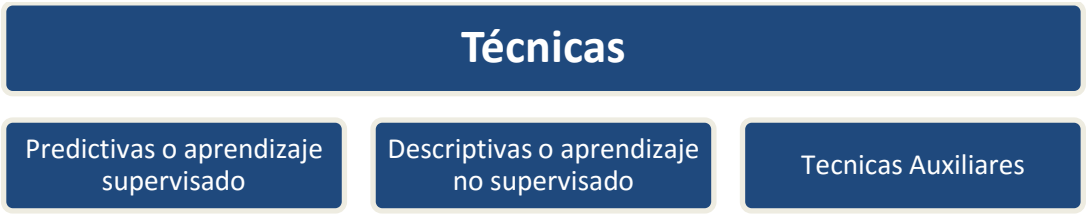

Figura 7. Clasificación de las técnicas de minería de datos.
Fuente: *(Santín Gonzalez & Peréz Lopéz, 2008)*

Las **Técnicas Predictivas** especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minado de datos antes de aceptarlo como válido. Formalmente, la aplicación de todo modelo tiene las siguientes fases:

Identificación objetiva, a partir de los datos se aplican reglas que permitan identificar el mejor modelo posible que ajuste los datos.

Estimación, proceso de cálculo de los parámetros del modelo elegido para los datos en la fase de identificación.

Diagnosis, proceso de contraste de la validez del modelo estimado.

Predicción, proceso de utilización del modelo identificado, estimado y validado para predecir valores futuros de las variables dependientes.

En algunos casos, el modelo se obtiene como mezcla del conocimiento obtenido antes y después del Data Mining (minería de datos) y también debe contrastarse antes de aceptar como válido. Por ejemplo. Las redes neuronales permiten descubrir modelos complejos y añadirlos a medida que progresa la exploración de los datos. Gracias a su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa. Podemos incluir entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. Tanto los árboles de decisión, como las redes neuronales y el análisis discriminante son a su vez técnicas de clasificación que pueden extraer perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato. Los árboles de decisión permiten clasificar los datos en grupos basados en los valores de las variables. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas (Santín Gonzalez & Peréz Lopéz, 2008).

En las **Técnicas Descriptivas** no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean dinámicamente partiendo del reconocimiento de patrones. En este grupo se incluyen las técnicas de clustering y segmentación (que también son técnicas de clasificación en cierto modo), las técnicas de asociación y dependencia, las técnicas de análisis exploratorio de datos y las técnicas de reducción de la dimensión (factorial, componentes principales, correspondencias, etc.) y de escalamiento multidimensional (Santín Gonzalez & Peréz Lopéz, 2008).

Tanto las técnicas predictivas como las técnicas descriptivas están enfocadas al descubrimiento del conocimiento embebido en los datos.

Las **Técnicas Auxiliares** son herramientas de apoyo más superficiales y limitadas, se trata de nuevos métodos basados en técnicas estadístico-descriptivas, consultas e informes y enfocados en general hacia la verificación (Santín Gonzalez & Peréz Lopéz, 2008).

A continuación, se visualiza en la Figura 8, la clasificación de las técnicas de minería de datos (Data Mining).

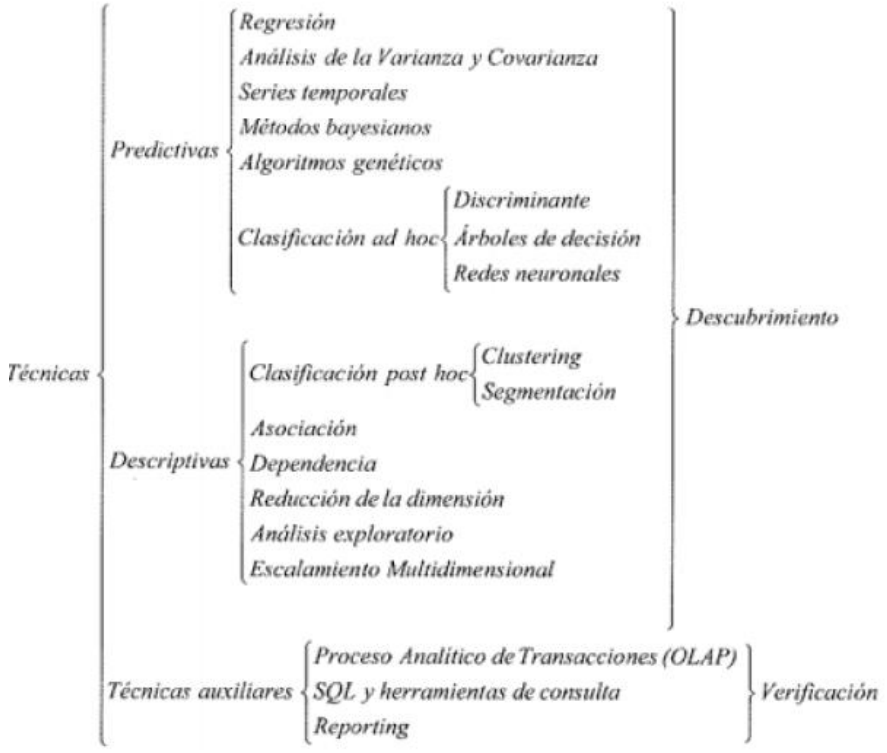

Figura 8. Clasificación de las técnicas de Data Mining
Fuente: *(Santín Gonzalez & Peréz Lopéz, 2008)*

2.1.1. Análisis y selección de la técnica de minería de datos

Según la investigación realizada y centrado en el propósito del presente estudio se utilizarán las técnicas predictivas o de aprendizaje supervisado, debido a que se basan en entrenar un modelo por medio de diferentes datos para poder predecir una variable partiendo de estos mismos datos.

2.2 Estudios de algoritmos analíticos

Un algoritmo en minería de datos es un conjunto de heurísticas y cálculos que permiten crear un modelo a partir de datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis en un gran número de iteraciones para determinar los parámetros óptimos para crear el modelo de minería de datos (Microsoft, 2018).

Las técnicas predictivas se aplican mediante algoritmos probados e implementados en soluciones de minería de datos, algunos de estos algoritmos son:

Algoritmos de Regresión: La Regresión Lineal Múltiple es un método matemático utilizado para crear un modelo de relación entre una variable dependiente, un número finito de variables independientes y una constante. La primera forma de regresiones lineales documentada fue el método de los mínimos cuadrados, el cual fue publicado por el matemático francés Adrien-Marie Legendre en 1805 y en dónde se incluía una versión del teorema de Gauss-Márkov.

En el campo de la educación, se ha utilizado el método de Regresión Lineal enfocado en aspectos físicos, sociológicos y psicológicos. Pero tanto en México como en otros países hay pocos estudios que relacionen la Trayectoria Escolar y el Rendimiento Académico apoyado en él. Por el tipo de seguimiento corresponde a un estudio longitudinal con cortes transversales. Por la recolección de la información es retrolectivo (cuando la obtención de la información es realizada una vez que la maniobra y el resultado han ocurrido) y por la direccionalidad de las observaciones es prolectivo (cuando la obtención de la información se realiza simultáneamente con la ocurrencia de la maniobra y, por lo tanto, simultáneo a la ocurrencia del resultado). Finalmente, por el análisis de la información y el tipo de estadísticos empleados, el estudio es analítico (Zaldivar et al., 2011).

Análisis de la varianza y covarianza: Este método es entender y determinar las similitudes entre dos o más grupos y generar conclusiones de igualdad. Una aproximación simple podría ser comparar las medias de los grupos y concluir sobre su igualdad o diferencia, aun así, el no tener en cuenta la varianza de los grupos de datos llevaría a conclusiones erróneas sobre el comportamiento de los mismos. Por lo tanto, se desarrolla en el análisis de datos de diseños experimentales el análisis de varianzas (ANOVA) y el análisis de covarianza (ANCOVA) como técnicas estadísticas que le permite al investigador hacer comparaciones entre medias de grupos observados.

Series temporales: Se define como un conjunto de observaciones de variables recogidas secuencialmente en el tiempo. Estos conjuntos se suelen recoger en instantes de tiempo equiespaciados. Si los datos se recogen en instantes temporales de forma continua, se debe o bien digitalizar la serie, es decir, recoger sólo los valores en instantes de tiempo equiespaciados, o bien acumular los valores sobre intervalos de tiempo (halweb, 2013).

Métodos bayesianos: Se dedica al estudio de algunos aspectos dentro de la inferencia estadística bayesiana, más concretamente, a los fundamentos de la inferencia bayesiana, las redes bayesianas y a distribuciones multivariantes específicas, como la distribución potencial exponencial (UCM, 2018).

Algoritmos genéticos: Los algoritmos genéticos (GA) son algoritmos de búsqueda y optimización basados en los mecanismos de selección natural y genética (Arredondo Vidal, 2009).

Están basados en el proceso genético de los organismos vivos. A lo largo de las generaciones, las poblaciones evolucionan en la naturaleza acorde con los principios de la selección natural y la supervivencia de los más fuertes, postulados por Darwin (1859). Por imitación de este proceso, los Algoritmos Genéticos son capaces de ir creando soluciones para problemas del mundo real. La

evolución de dichas soluciones hacia valores óptimos del problema depende en gran medida de una adecuada codificación de las mismas (Arredondo Vidal, 2009).

A continuación, en la Figura 9 se describen las técnicas predictivas relacionadas con los algoritmos correspondientes.

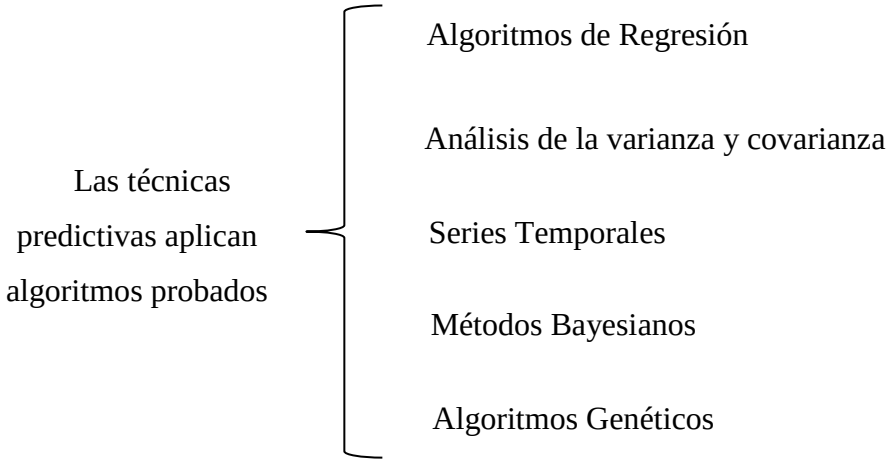

Figura 9. Clasificación de Algoritmos de Data Mining

KDnuggets es un sitio líder en el tema de Minería de datos y está editado por Gregory Piatetsky-Shapiro considerado uno de los máximos exponentes de analítica a nivel mundial. Actualmente, este sitio llega a más de 500.000 visitantes mensuales únicos y más de 200.000 suscriptores a través de correo electrónico, Twitter, LinkedIn, Facebook, feedly / RSS y Google+ *(Piatetsky-Shapiro, 2018)*.

La última encuesta realizada por KDnuggets a 844 profesionales permitió identificar la lista de los principales algoritmos utilizados por Data Scientists, en donde la pregunta era la siguiente.

¿Qué métodos / algoritmos usó en los últimos 12 meses para una aplicación real relacionada con la Ciencia de Datos?, la encuesta desplego los siguientes resultados visualizados en la Figura 10.

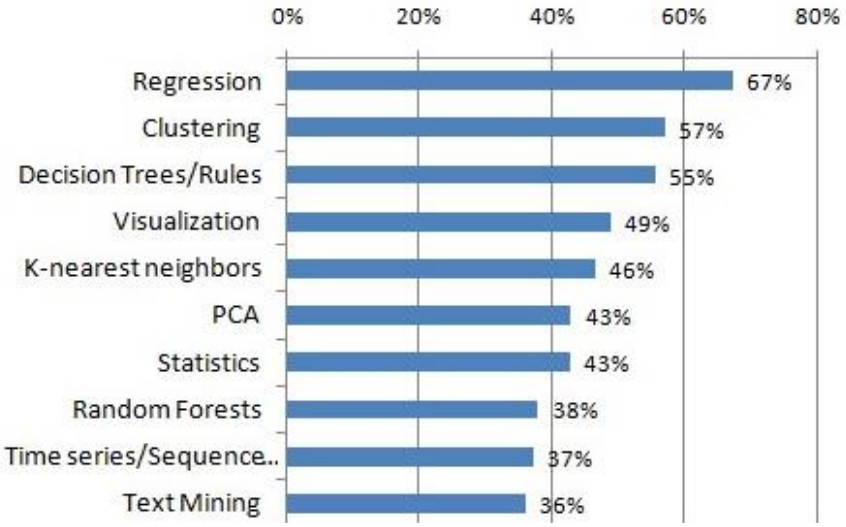

Figura 10. Los 10 algoritmos y métodos más utilizados por los científicos de datos.
Fuente: *(Gregory Piatetsky, 2016)*

Nota: Estadísticas o Visualización (y varias otras opciones) no son algoritmos, pero se pueden describir mejor como métodos o enfoques, por tal motivo la gráfica representa a "Los 10 algoritmos y métodos principales utilizados por los científicos de datos".

De la misma manera, la encuesta ejecutada por KDnuggets realizó un análisis de la utilización de los algoritmos y métodos por tipo de empleo, en donde se nota que casi todos usan algoritmos de aprendizaje supervisado, como se visualiza en la Tabla 3.

Tabla 3

Los 10 principales algoritmos y métodos utilizados por tipo de empleo

Algorithm	Industry	Government/Non-profit	Academia	Student	All
Regression	71%	63%	51%	64%	67%
Clustering	58%	63%	51%	58%	57%
Decision	59%	63%	38%	57%	55%
Visualization	55%	71%	28%	47%	49%
K-NN	46%	54%	48%	47%	46%
PCA	43%	57%	48%	40%	43%
Statistics	47%	49%	37%	36%	43%
Random Forests	40%	40%	29%	36%	38%
Time series	42%	54%	26%	24%	37%
Text Mining	36%	40%	33%	38%	36%
Deep Learning	18%	9%	24%	19%	19%

Fuente: *(Gregory Piatetsky, 2016)*

2.2.1 Análisis y selección del algoritmo

De acuerdo a la investigación realizada podemos mencionar lo siguiente:

- Observamos que los científicos de datos a nivel de industria son más propensos a usar algoritmos de Regresión, Decision, Clustering y Visualización.
- Es más probable que el gobierno utilice Visualizacion, Regresion y Clustering.
- Los investigadores académicos y los estudiantes son más propensos a usar Regresión y Clustering (Gregory Piatetsky, 2016).

Según este análisis realizado en los diferentes campos de la utilización de los algoritmos analíticos, podemos observar que los más utilizados son Regresión y Clustering, según lo menciona estudios e investigaciones realizadas referente a Business Analytics, Big Data, Data Mining, Data Science y Machine Learning a nivel mundial.

Adicionalmente, de acuerdo a una investigación de implementaciones a nivel mundial se pudo evidenciar que grandes empresas como Ford Motor Company (Blacke & Jacobson, 2016),

Automotores y Anexos S.A., AYASA (itahora, 2018), Seguros Equinoccial (IT Ahora, Seguros: La forma de hacer negocios está cambiando, 2018), Yanbal International (IT Ahora, 2018), Wester Union (Bulkley, 2018) y Adidas (Underwood, 2017) están implementando proyectos de analítica avanzada, relacionados al conocimiento de la demanda de sus organizaciones, utilizando algoritmos de regresión y time series.

Los datos son el activo principal de estas compañías, por tal motivo han optado por formar un equipo centralizado de ciencia de datos organizado para compartir análisis, mejores prácticas y difundir información optimizada, basada en datos para la toma de decisiones en toda la organización.

De acuerdo a lo analizado en el mercado de la utilización de los algoritmos analíticos, se pudo evidenciar que los algoritmos más utilizados por los científicos de datos en implementaciones de analítica avanzada son Regresión y Clustering lo que constituyen nuestro primer filtro en el presente trabajo de investigación, por tal motivo se procede a realizar un análisis a nivel de dos parametrizaciones: implementaciones a nivel de gobierno y casos de éxito en empresas a nivel mundial, como se visualiza en la Tabla 4.

Tabla 4

Análisis de los algoritmos

ALGORITMO	IMPLEMENTACIONES A NIVEL DE GOBIERNO	CASOS DE ÉXITO EN EMPRESAS A NIVEL MUNDIAL
REGRESION	✓	✓
CLUSTERING	✓	X

Como una forma de obtener resultados totales se ha generado en la Figura 11 un resumen de la valoración obtenida en la Tabla 4, lo que deduce según la investigación realizada que la utilización con más éxito y favoritismo en proyectos de analítica avanzada específicamente para predicciones es el algoritmo de **Regresión**, la misma que será aplicada en el presente trabajo de investigación.

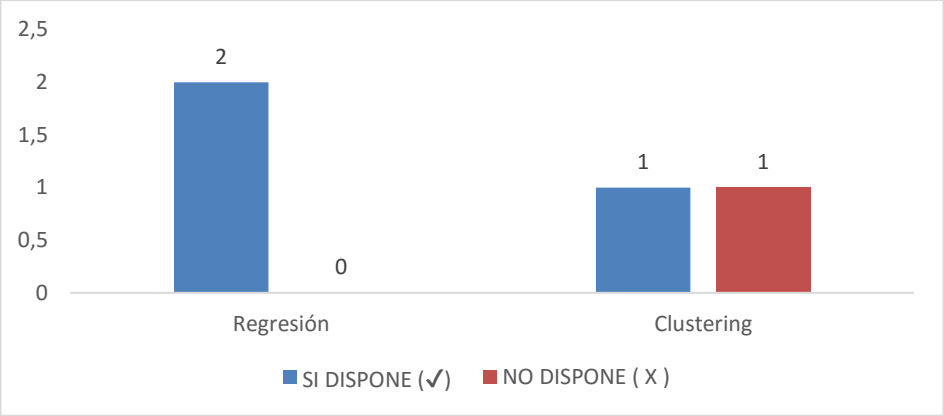

Figura 11. Resultados de la evaluación de los algoritmos analíticos

2.3 Estudios realizados por empresas especializadas

Debido a factores externos, como la gran diversidad de herramientas y avances de la tecnología, se considera como una buena alternativa revisar estudios realizados por empresas especializadas en investigar fortalezas y debilidades de los productos, soluciones y herramientas, así como las tendencias del mercado. Por lo anterior, se menciona el análisis realizado por las empresas; Gartner y Forrester Wave™.

Gartner es una empresa de consultoría e investigación del mercado de las nuevas tecnologías dedicadas exclusivamente a investigar y analizar las tendencias del mercado, para posteriormente y basándose en las conclusiones, elaborar un ranking de los fabricantes con las mejores soluciones y productos. Gartner incluye entre sus clientes a algunas de las más grandes empresas, agencias de gobierno, empresas tecnológicas y agencias de inversión como BT, CV, The Wall Street Journal, etc.

La empresa se concentra en la investigación, programas ejecutivos, consultas y eventos. Los resultados de sus estudios son presentados bajo el nombre de “cuadrante mágico de Gartner” (SIAG Consulting, 2016).

La interpretación de los cuadrantes de Gartner es la siguiente:

El primer cuadrante - **Líderes**, donde se encuentran los proveedores que mayor puntuación han obtenido como resultado de combinar su gran capacidad de visión del mercado y la habilidad para ejecutar.

El segundo cuadrante - **Visionarios,** semejantes a los líderes en su capacidad para anticiparse a las necesidades del mercado, pero no disponen de medios suficientes para realizar implantaciones globales.

El tercer cuadrante - **Retadores o aspirantes**, ofrecen buenas funcionalidades, pero tienen menor variedad de productos al estar centrados en un único aspecto de la demanda del mercado.

El cuarto cuadrante - **Jugadores de nicho**, no llegan a puntuar lo suficiente en ninguna de las dos categorías (SIAG Consulting, 2016).

El Cuadrante Mágico Gartner 2018 para Plataformas de Ciencia de Datos evaluó 16 firmas de análisis sobre múltiples criterios y los colocó en 4 cuadrantes, en función de la amplitud de la visión y la capacidad de ejecución.

El análisis se puede visualizar en la Figura 12, en la cual los líderes del mercado son: KNIME, Alteryx, SAS, RapidMiner, H2O.ai

Figura 12. Cuadrante Mágico de Gartner para plataformas de ciencia de datos
Fuente: *(Gartner, 2018)*

Del Cuadrante Mágico de Gartner se han obtenido las empresas categorizadas de la siguiente manera:

Primer cuadrante - Líderes (5): KNIME, Alteryx, SAS, RapidMiner, H2O.ai

Segundo cuadrante - Challengers (2): MathWorks, TIBCO Software.

Tercer cuadrante - Visionarios (5): IBM, Microsoft, Domino Data Lab, Dataiku, Databricks.

Cuarto cuadrante - Jugadores de Nicho (4): SAP, Angoss, Anaconda, Teradata

Por otro lado, **Forrester**, como empresa evaluadora es considerada como una de las firmas de investigación y asesoramiento sin fines de lucro más influyentes del mundo, su investigación se rige en el análisis de las tendencias del mercado utilizando personal especializado en torno al tema.

El conocimiento obtenido por Forrester se basa en encuestas anuales de más de 675,000 consumidores y líderes de negocios en todo el mundo, mediante el uso de metodologías rigurosas y objetivas, así como la sabiduría compartida de nuestros clientes más innovadores. (Forrester, 2018)

El último informe de Forrester sobre análisis predictivo (Predictive Analytics) y soluciones de aprendizaje automático (Machine Learning Solutions) fue publicado en el primer trimestre de 2017, escrito por Mike Gualtieri. En el mismo se examina y evalúa a 14 empresas en términos de estrategia, oferta actual y presencia en el mercado.

Los resultados se resumen en la Figura 13, en donde lideran el mercado SAS, IBM, SAP, Rapidminer, Knime, Angoss y FICO.

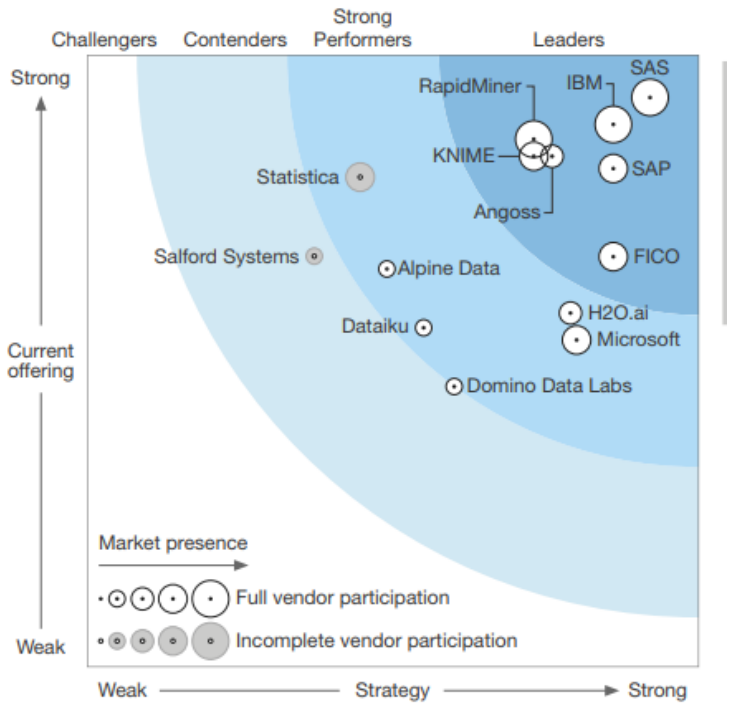

Figura 13. Forrester Wave™: Predictive Analytics and Machine Learning Solutions, Q1 2017
Fuente: *(Gregory Piatetsky, 2017)*

Gartner y Forrester usan diferentes metodologías, sin embargo en ambos casos, cuanto más lejos esté el círculo que representa a una empresa de la esquina inferior izquierda del diagrama, es mejor.

La evaluación realizada por Gregory Piatetsky-Shapiro uno de los principales influyentes de Big Data, Data Mining y Data Science, sobre Gartner vs Forrester de Data Science, Predictive Analytics y Machine Learning durante el primer trimestre del 2017, despliega como líderes indiscutibles a SAS, IBM, Rapidminer y Knime, cuyo detalle se visualiza en la Figura 14.

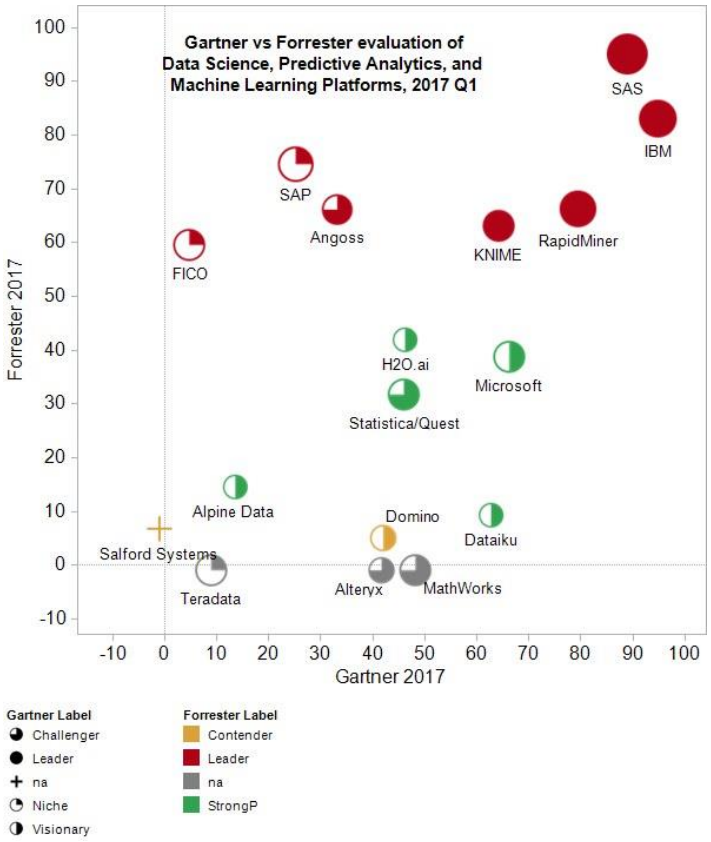

Figura 14. Evaluación de Gartner vs Forrester de Data Science, Predictive Analytics y Machine Learning Platforms, 2017 Q1
Fuente: *(Piatetsky, 2018)*

El tamaño del círculo corresponde al tamaño estimado del proveedor, el color es *Forrester Label* y la forma (cómo está lleno el círculo) es *Gartner Label.* En total, hay 17 empresas: 13 que aparecen en ambas, 3 solo en Gartner (abreviado como G), y una solo en Forrester (abreviada como F).

De la figura anterior vemos varios clusters que a continuación se detallan:

Líderes fuertes: SAS, IBM, RapidMiner y KNIME están clasificados como líderes tanto por G & F.

Líderes: Angoss, SAP y FICO son líderes de F pero solo Nichos / Challenger para G.

Grandes competidores: H2O.ai, Microsoft y Statistica / Quest son buenos intérpretes para F y Visionary / Challenger para G

Contendientes: Alpine Data, Domino, Dataiku: fuertes intérpretes / contendientes en F, visionarios en G

Jugadores: Salford Systems, Teradata, Alteryx, MathWorks: solo tienen una clasificación

A continuación, se analizan las firmas **líderes** según lo mencionan Gartner y Forrester en sus publicaciones del 2017. Detallando una breve descripción en cada caso:

KNIME proporciona la plataforma de código abierto KNIME Analytics, con más de 100.000 usuarios en todo el mundo. KNIME ofrece una amplia red de colaboración y de aprendizaje online. En 2017, KNIME agregó versiones en la nube de su plataforma para AWS y Microsoft Azure, (Piatetsky, 2018)

SAS proporciona muchos productos de software para análisis y ciencia de datos. Para este nuevo informe Gartner evaluó SAS Enterprise Miner (EM) y el paquete de productos SAS Visual Analytics (Piatetsky, 2018).

Gartner menciona que:

"…SAS sigue siendo un Líder, pero ha perdido algo de terreno en términos de integridad de la visión y capacidad de ejecución. El paquete de Visual Analytics es prometedor debido a su arquitectura lista para la nube Viya, que es más abierta que la arquitectura SAS anterior y hace que

la analítica sea más accesible para una amplia gama de usuarios. Sin embargo, un enfoque confuso de multiproducto ha empeorado la completitud de la visión de SAS, y la percepción de los altos costos de licenciamiento ha perjudicado su capacidad de ejecución. A medida que el enfoque del mercado se traslada al software de código abierto y la flexibilidad, la lentitud de SAS para ofrecer una plataforma abierta y cohesiva ha pasado factura…" (Gartner, 2018).

RapidMiner incluye la herramienta de desarrollo de modelos RapidMiner Studio (disponible tanto en edición libre como comercial), RapidMiner Server y RapidMiner Radoop (Piatetsky, 2018).

Gartner menciona que:

"…RapidMiner sigue siendo un líder al ofrecer una plataforma completa y fácil de usar para todo el espectro de científicos de datos y equipos de ciencia de datos. RapidMiner continúa enfatizando la ciencia de datos básicos y la velocidad de desarrollo y ejecución del modelo mediante la introducción de nuevas capacidades de productividad y rendimiento…" (Gartner, 2018)

IBM proporciona muchas soluciones analíticas. Para este MQ, Gartner evaluó SPSS Modeler y SPSS Statistics, pero no Data Science Experience (DSX), que no cumplía los criterios de Gartner para la evaluación del eje Ability to Execute (Piatetsky, 2018).

Gartner escribe:

"…IBM ahora es un visionario, habiendo perdido terreno en términos de integridad de visión y capacidad de ejecución, en relación con otros proveedores. La oferta DSX de IBM, sin embargo, tiene el potencial de inspirar una visión más integral e innovadora. IBM ha anunciado planes para ofrecer una nueva interfaz para sus productos SPSS en 2018, una que integra completamente SPSS Modeler en DSX…" (Gartner, 2018).

2.3.1 Análisis y selección de la herramienta analítica

De acuerdo a la investigación realizada de herramientas analíticas se puede mencionar a: **KNIME, SAS y RapidMiner** como las plataformas mejor posesionadas en el mercado, esta clasificación es realizada por empresas reconocidas a nivel mundial, y especializadas en investigar las fortalezas y debilidades de las plataformas tecnológicas como Gartner y Forrester.

Para la selección de las herramientas se formó un grupo focal conformado por 4 profesionales de Business Analytics, para la clasificación de los participantes se utilizó la técnica de muestra de conveniencia, esto debido a que los individuos fueron seleccionados porque están a la mano o son fáciles de encontrar. Estos profesionales no tienen sesgo alguno por las herramientas analizadas, los mismos son consultores independientes que han realizado varios proyectos entorno a inteligencia de negocios y analítica avanzada, como se visualiza en la Tabla 5.

Tabla 5

Participantes del Focus Group

Integrantes:	**Rol**	**Perfil Profesional**	**Proyectos Realizados**
María Ortuño	Moderador	Ing. en Sistemas Informáticos e investigador en el presente trabajo de investigación.	Investigador del presente trabajado de investigación
Karina Mazón	Participante	Ing. en Sistemas Informáticos, con 5 años de experiencia realizando proyectos de BI, con implementaciones realizadas en las empresas más reconocidas a nivel nacional (LinkedIn_KM, 2018).	- DISS (Directorio de Servicios Sociales) con módulo BI. - PACHA (aplicación móvil creada para agricultores). - Proyecto de Big Data en Corporación favorita. - Proyecto de BI en Vallejo Araujo. - Proyecto de analítica en CONFITECA C.A.
Yoann Leny	Participante	Más de 8 años de experiencia realizando proyectos de Big Data, inteligencia de negocios y analítica	- Proyecto en INTELLIPHARM INC (solución industrial para un grupo de farmacias y fabricantes para analizar el rendimiento en las tiendas.).

CONTINÚA

		avanzada con implementaciones realizadas alrededor del mundo entre ellas en Australia, Francia y Estados Unidos (LinkedIn_YL, 2018).	- Proyectos en SmartMorph utilizando herramientas de BI y analítica avanzada.
Hugo Vera Flores	Participante	Master en Gerencia en Sistemas y Arquitectura Empresarial, Consultor senior en Data Warehouse y Business Intelligence, certificado SAP Business Objects Profesional ,SAP Business Objects Data Services, Tableau Software Certify, Emprendedor y fundador de Business Information Solutions S.A empresa dedicada a proveer soluciones de Business Intelligence, Data Warehouse, EPM y Big Data (LinkedIn_HV, 2018).	- Proyecto Business Intelligence - Eni Ecuador-AGIP GAS - Proyecto BI en Corporación GPF - Farcomed - Proyecto BI en Rosa Prima - SAP Strategy Management - TIOSA - Data Mart CRM - Corporación Favorita - TLOG - CRM Project - Corporacion Favorita - Business Intelligence Project Andalucia Coop OBI - Sistema de Información Empresarial - PETROPRODUCCION - Business Intelligence Project - Seguros Colonial.
Jose Alvarado	Participante	Consultor senior con más de 16 años de experiencia en proyectos de Business Intelligence y Analítica Avanzada a nivel nacional, con experiencia trabajando con herramientas como Tableau Software, Knime, SAP y Alteryx (LinkedIn_JA, 2018).	- Proyecto BI en Corporación GPF. - Proyecto BI en Vallejo Araujo. - Proyecto BI en Corporación Favorita. - Proyecto BI en Automotores y Anexos. - Proyecto BI en General Motors.

Las herramientas analíticas a ser evaluadas son SAS Enterprise Miner (licencia), RapidMiner (free) y Knime (free), estas debido a su presencia en el mercado según lo investigado y analizado. Con los integrantes del Focus group se ha evaluado las características más importantes de cada herramienta, el producto de la evaluación se visualiza en la Figura 15.

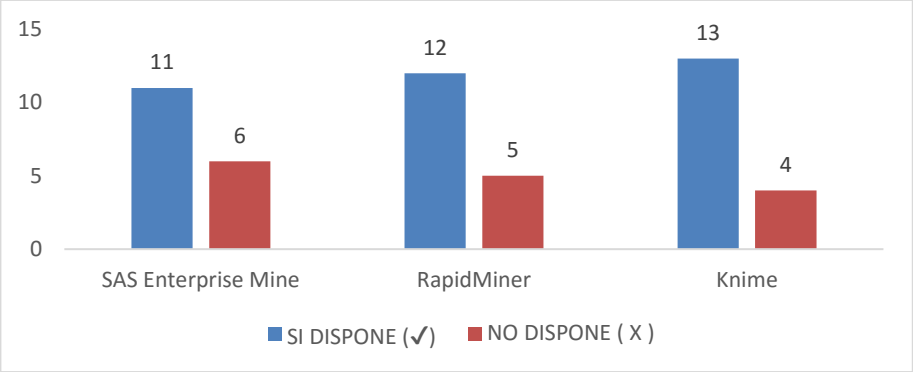

Figura 15. Resultados de evaluación de las herramientas analíticas

A continuación, se detallan las características de la evaluación de cada una de las herramientas analíticas, como se visualiza en la Tabla 6.

Tabla 6

Detalle de evaluación de las herramientas analíticas

	CARACTERÍSTICAS	**HERRAMIENTAS**		
		SAS Enterprise Mine	**RapidMiner**	**Knime**
1	Licencia libre	X	✓	✓
2	Multiplataforma	✓	✓	✓
3	Puede combinar modelos	✓	✓	✓
4	Interfaz amigable	✓	✓	✓
5	Permite visualización de datos	✓	✓	✓
6	Flexibilidad	X	✓	✓
7	Fácil de Configurar	X	✓	✓
8	Fácil de Instalar	✓	✓	✓
9	Conversión fácil de datos	✓	✓	✓
10	Filtros	✓	✓	✓
11	Integración con R	X	✓	✓
12	Personalización de módulos de predicción incorporados	X	X	X
13	Creación de macros, módulos y la posibilidad de la reutilización de los mismos.	X	X	✓
14	Procesamiento de grandes volúmenes de información	✓	X	X

15	Soporte en Ecuador, Partners según lo enuncian sus páginas oficiales	✓	X	X
16	Casos de éxito expuestos en sus páginas oficiales	✓	X	X
17	Validación del modelo	✓	✓	✓

La evaluación realizada de las herramientas analíticas por el focus group fue expuesta y analizada por el CNE, en la cual en conjunto con el investigador del presente trabajo de investigación se decidió utilizar la herramienta open source **Knime Analitycs Platform**, debido a que contiene las características más importantes para un proyecto exitoso de analítica avanzada, además la herramienta mencionada contiene algoritmos de predicción pre-configurados, entre ellos el modelo de Regresión, permitiendo a los analistas la utilización de los mismos sin necesidad de ser expertos en estadística; así también se consideró que la institución donde se está aplicando dicha investigación pertenece al sector público y se acogió el decreto ejecutivo 135 de normas de optimización y austeridad del gasto público dispuesto por el Presidente Constitucional de la República del Ecuador.

En vista de los argumentos presentados e investigados la técnica, algoritmo y herramienta seleccionada son: técnica predictiva, algoritmo de Regresión y la herramienta analítica Knime Analitycs Platform; las mismas que serán aplicadas en el presente trabajo de investigación.

CAPITULO III

PROPUESTA DE UN MODELO ANALÍTICO PARA LA PREDICCIÓN DEL PRESUPUESTO ELECTORAL CNE ECUADOR

En esta sección, se especifica la propuesta del modelo analítico para la predicción del presupuesto electoral, utilizando la técnica, algoritmo y herramienta seleccionada en el capítulo anterior.

3.1 Proceso actual de toma de decisiones

Actualmente el proceso de proyección del presupuesto se lo realiza manualmente mediante cálculos en excel y revisiones de presupuestos en elecciones anteriores, debido a que no existe un software que sustente los valores emitidos, a continuación, se explica el proceso de las elecciones generales del 19 de febrero de 2017, en donde se realizó las elecciones de presidente, asambleístas nacionales, asambleístas provinciales, asambleístas del exterior y parlamentarios andinos.

El Registro Civil envió una carga inicial al Consejo Nacional Electoral(CNE) del registro de los ciudadanos, este proceso se lo realizó en febrero de 2016, luego la Dirección Nacional de Sistemas e Informática Electoral realiza la depuración de la información para obtener la base del número de electores que existen a la fecha, posterior aquello las actualizaciones de información de los electores se realizan cada 3 meses y luego cada mes, de manera de tener actualizado la base de datos; el Consejo Nacional Electoral CNE emite al Ministerio de Economía y Finanzas el consolidado de la proyección del presupuestos para las próximas elecciones de manera de poder tener capital para realizar diferentes gastos entre ellas impresión de papeletas, contratación de personal, adquisición de equipos, etc., todo lo que involucra un proceso electoral, centrándonos en la Dirección de Logística emite su presupuesto proyectado pero agrega un gasto adicional del 30% de electores, esto debido a que considera que la población de electores crece cada día pero como

no sabe en qué porcentaje crecerá de acuerdo al conocimiento y experiencia en otro procesos electorales asume el crecimiento de ese valor.

El Ministerio de Economía y Finanzas emite un comunicado al CNE del reajuste del presupuesto debido a que el mismo está sobredimensionado y no existen fondos para tal gasto, una vez que existe un acuerdo de presupuesto el dinero es emitido a la entidad emisora. El registro electoral se cerró en noviembre de 2016 en donde se visualizó que la población de electores no creció un 30% sino solamente un 15% lo que provoco que luego de las elecciones el dinero sobredimensionado sea devuelto al Ministerio de Economía y Finanzas provocando que la diferencia devuelta no pueda ser utilizada por otras direcciones del CNE que en verdad necesitan del mismo, como se visualiza en la Figura 16.

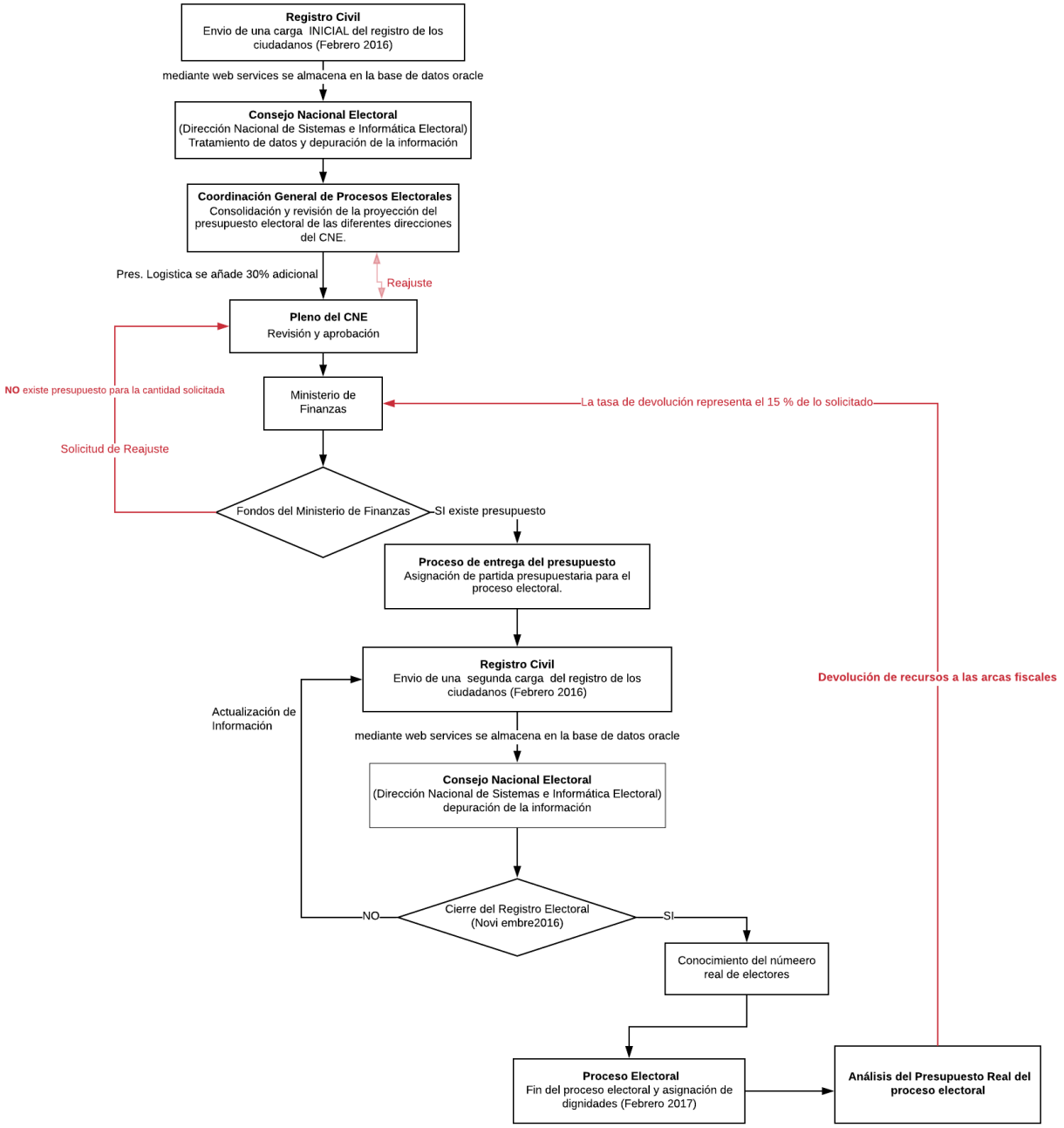

Figura 16. Diagrama de flujo Elecciones Generales 19 febrero de 2017

Fuente: Ing. Laura Ayala – Responsable de la Dirección Nacional de Logística de CNE

A continuación, en la Figura 17 se detalla el diagrama de procesos.

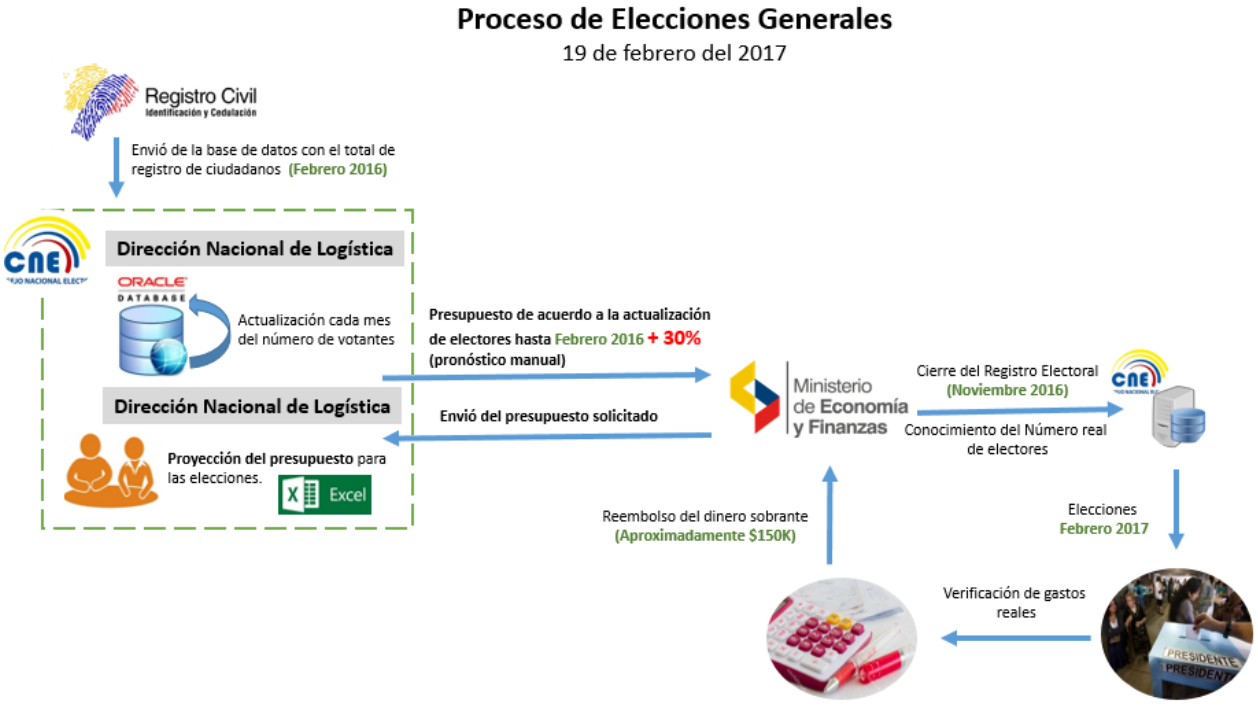

Figura 17. Descripción de Proceso que realizó cada entidad participante en las elecciones del 2017

Fuente: Ing. Laura Ayala – Responsable de la Dirección Nacional de Logística de CNE

3.2 Desarrollo de la propuesta

De acuerdo al análisis del negocio de la Dirección Nacional de Logística de CNE y la necesidad de poder automatizar la proyección del presupuesto electoral se ha propuesto la arquitectura más idónea y fácil para cubrir los requerimientos del área.

Debido a cuestiones de centralización de información y análisis, en conjunto con el área de Logística de CNE se ha tomado la decisión de incluir el proceso de proyección de la demanda a la solución de inteligencia de negocios, realizando mejoras y recomendaciones en cada una de las capas.

Aunque se pueden encontrar muchas variantes y elementos dentro de las arquitecturas de inteligencia de negocios implementadas en las empresas, la mayor parte de ellas incluyen una serie de componentes principales. A continuación, se visualiza en la Figura 18 las diversas capas y componentes de cada una de ellas las cuales debe existir en la arquitectura de los sistemas de BI.

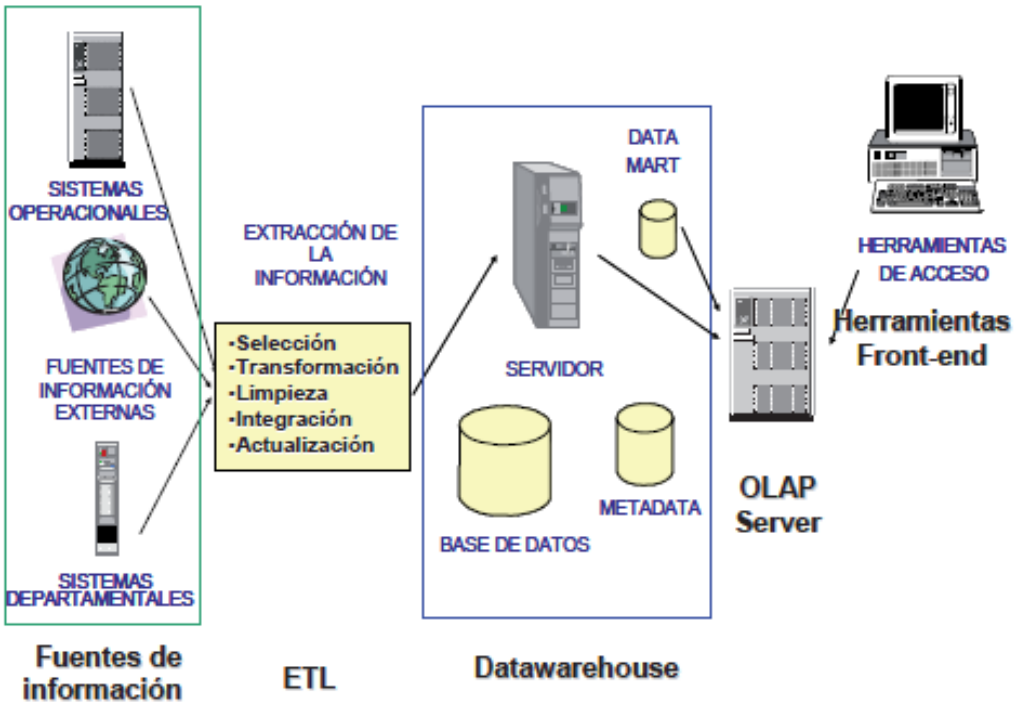

Figura 18. Componentes de Business Intelligence

Fuente: *(Cano, 2007)*

Los componentes son:

- **Fuentes de información**, generalmente son bases de datos relacionales, hojas de excel, archivos planos que almacenan la información transaccional de la organización y en ocasiones datos complementarios o externos (hfrancob, 2011), de las cuales partiremos para alimentar de información al DatawareHouse.
- **Proceso ETL (Extracción, Transformación y Carga)**, antes de almacenar los datos en un datawarehouse, éstos deben ser transformados, limpiados, filtrados y redefinidos.

Normalmente, la información que tenemos en los sistemas transaccionales no está preparada para la toma de decisiones.

- **Datawarehouse o almacén de datos, con el Metadata o Diccionario de datos**, se busca almacenar los datos de una forma que maximice su flexibilidad, facilidad de acceso y administración.
- **El motor OLAP**, que nos debe proveer capacidad de cálculo, consultas, funciones de planeamiento, pronóstico y análisis de escenarios en grandes volúmenes de datos.
- **Las herramientas de visualización**, que nos permitirán el análisis y la navegación a través de los mismos.

Es fundamental definir todas las capas de la implementación de la solución BI (Business Intelligence), y sobre todo la parte en donde está involucrado el módulo de la proyección del presupuesto electoral para presentar a la institución todos los componentes de la implementación. Los componentes utilizados en el desarrollo de la propuesta de solución, se describen en la siguiente figura:

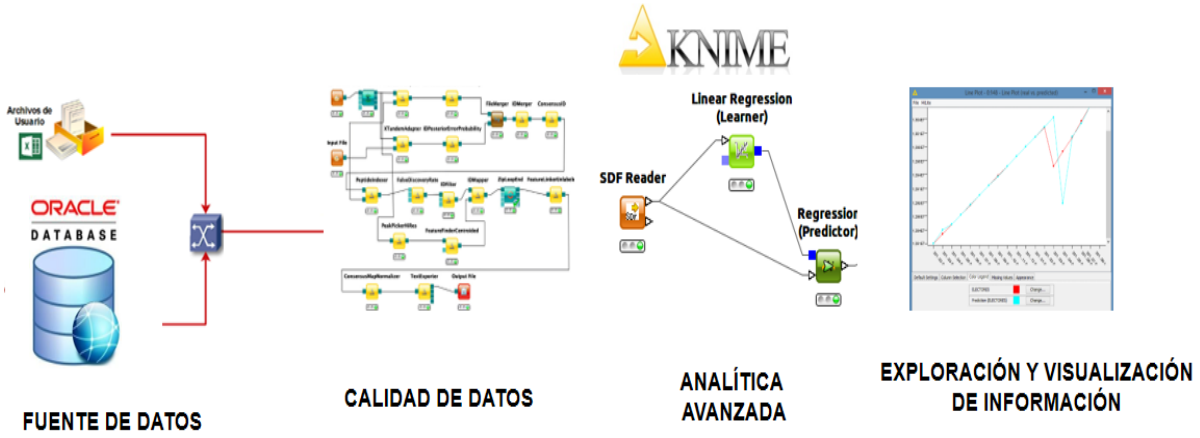

Figura 19. Componentes utilizados en el desarrollo de la propuesta

A continuación, se detalla cada uno de los componentes a utilizar para el desarrollo de la propuesta:

Fuentes de datos que son el núcleo de la solución, el CNE almacena toda la información recopilada de sus diferentes fuentes en la base de datos Oracle versión 12c, y desde la cual se realizó la extracción de la información necesaria.

Calidad de datos, la misma que se realizó con Knime Analytics Platform herramienta de Data Quality y analítica avanzada. Software de código abierto para crear aplicaciones y servicios de ciencia de datos, nuevos desarrollos intuitivos, abiertos y de integración continua, KNIME hace que la comprensión de datos y el diseño de flujos de trabajo de ciencia de datos y componentes reutilizables sean accesibles para todos (KNIME, 2018).

Cabe mencionar que, por el tamaño de información que maneja la institución no se recomienda la implementación aún de una base de datos analítica, debido a que la transaccionalidad que manejan es baja y solamente a demanda.

Minería de datos y analítica avanzada, esta etapa se realizó con la herramienta analítica Knime Analytics Platform debido a que a más de ser útil para data quality es un software potente para analítica avanzada y la mejor evaluada de acuerdo a la investigación realizada en el presente trabajo de investigación.

En esta fase se aplicó el algoritmo de **Regresión** para el proceso de proyección del presupuesto y cuyo modelo viene pre-configurado en la herramienta analítica Knime.

Exploración y visualización de la información, en la misma herramienta Knime se realizó todo el tema de análisis de la data, debido a que contiene componentes interactivos de visualización, el cual es un producto de análisis de datos que ayuda a mejorar y acelerar la toma de decisiones.

A pesar que Knime Analytics platform contiene un módulo de visualización se recomienda la utilización de una herramienta de inteligencia de negocios de manera de tener una gobernabilidad de la información dentro de la institución.

3.2.1 Desarrollo de la solución

Como se ha mencionado anteriormente, para el desarrollo de la investigación y para fines de análisis se utilizará la metodología CRISP-DM. Esta metodología es una de las más utilizadas en el área de minería de datos, que incluye un modelo y una guía estructurada en seis fases, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores. Cada etapa de la metodología CRISP-DM fue ajustada a la realidad de los procesos realizados en la investigación dentro del Consejo Nacional Electoral. A continuación, se describen las actividades que se realizó en cada fase:

FASE I: Comprensión del negocio

Esta primera fase es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Para obtener el mejor provecho de la minería de datos, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados (Chapman et al., 2000). Las actividades contempladas dentro de esta fase son: Determinar los objetivos del negocio, Evaluación de la situación, Determinar los objetivos de la minería de datos y realizar el plan del proyecto.

En referencia a los acuerdos planteados en la minuta de reunión de inicio del proyecto, visualizada en la Tabla 9 del Anexo 1, se realizó la proyección del presupuesto electoral en base a los siguientes criterios:

- Se aplicó la metodología CRISP-DM debido a que es la más robusta en la realización de proyectos de minería de datos.
- Se utilizó la técnica predictiva y el algoritmo de regresión para los procesos de proyección del presupuesto electoral, por ser los mejores valorados en el presente trabajo de investigación.
- En el proyecto analítico se utilizó la plataforma analítica KNIME 3.5.3 por ser la mejor evaluada en el presente trabajo de investigación.
- Se realizó la predicción del presupuesto electoral dentro de un periodo de 12 meses (1 año), debido a que el CNE antes de los 12 meses emite el presupuesto a utilizar al ministerio de Economía y Finanzas.

En conclusión, de acuerdo a la fase I, se realizó la reunión de inicio del proyecto para el conocimiento de los objetivos del negocio, evaluación de la situación actual, establecimiento de los objetivos de la minería de datos y la generación del plan del proyecto con la Ing. Laura Ayala , responsable de la Dirección Nacional de Logística de CNE, también se solicitó compromiso al personal que va a formar parte del proceso de implementación del proyecto, así como el tiempo para que puedan realizar los análisis de información.

FASE II: Comprensión de los datos

Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos (Chapman et al., 2000).

Las actividades realizadas en esta fase son:

a) **Recopilación inicial de datos:** Todos los datos necesarios para la generación del modelo se encuentran almacenados en una base de datos Oracle 12c, por tal motivo no hubo necesidad de aplicar técnicas de recolección de datos debido a que en Oracle está centralizado toda la información del negocio, a través de la herramienta analítica Knime se realizó la conexión directa con la base de datos utilizando el usuario de lectura otorgado por el área de sistemas del CNE, como se visualiza en la Figura 20.

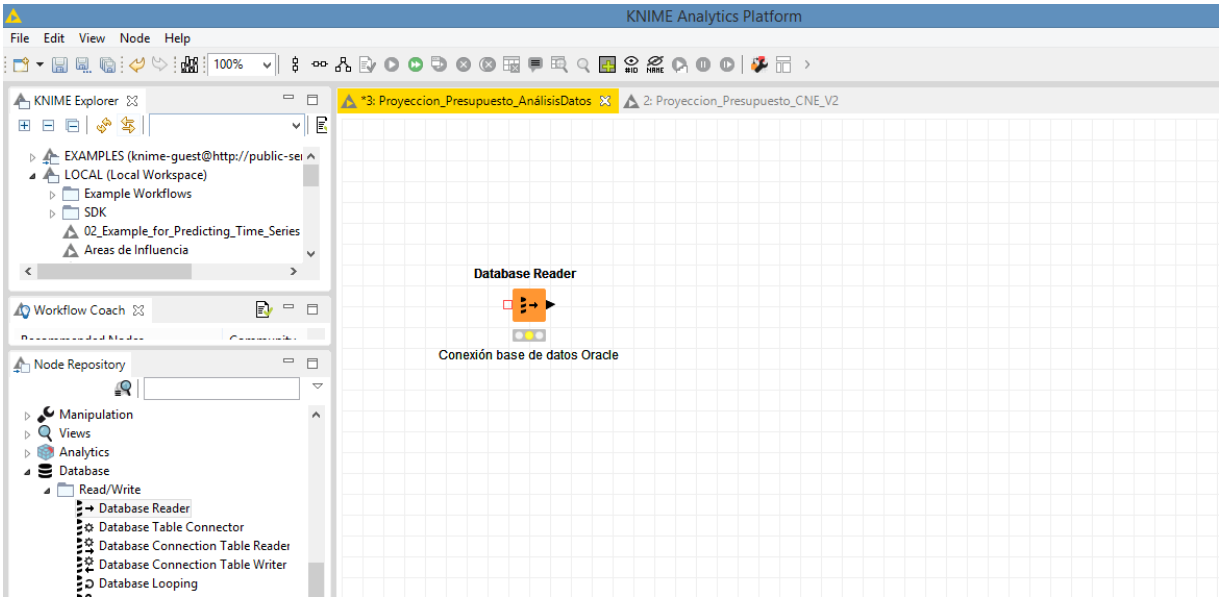

Figura 20. Conexión de Knime con la base de datos Oracle

b) **Descripción y exploración de los datos:** Después de obtener los datos iniciales, estos deben ser descritos. Este proceso implica establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

La herramienta analítica Knime incorpora componentes predefinidos para la exploración de la información, entre ellas **Statistics y Data Explorer**.

En la Figura 21, se visualiza el resultado del componente **Statistics** que principalmente visualiza medidas estadísticas como mínimo, máximo, promedio, desviación estándar, varianza, mediana, suma global, número de valores perdidos, nulos, etc; este componente me permite explorar el comportamiento de los datos.

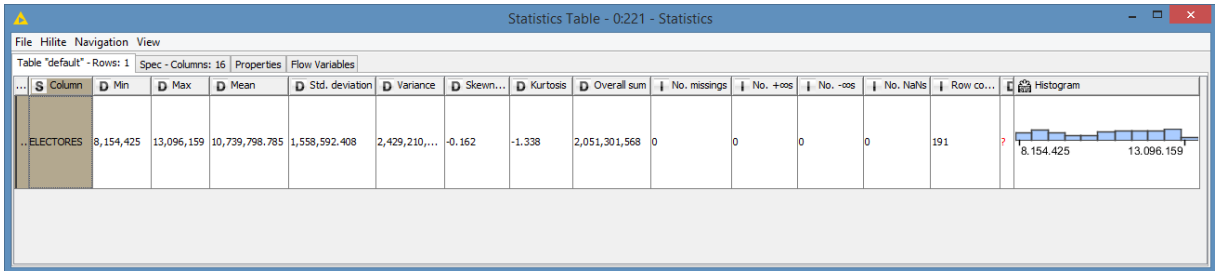

Figura 21. Estadísticas básicas del componente Statistics

Adicional en el mismo componente Statistics podemos visualizar el tipo de dato, como se visualiza en la Figura 22, la misma que permite tener una idea de la descripción de la información devuelta por la base de datos.

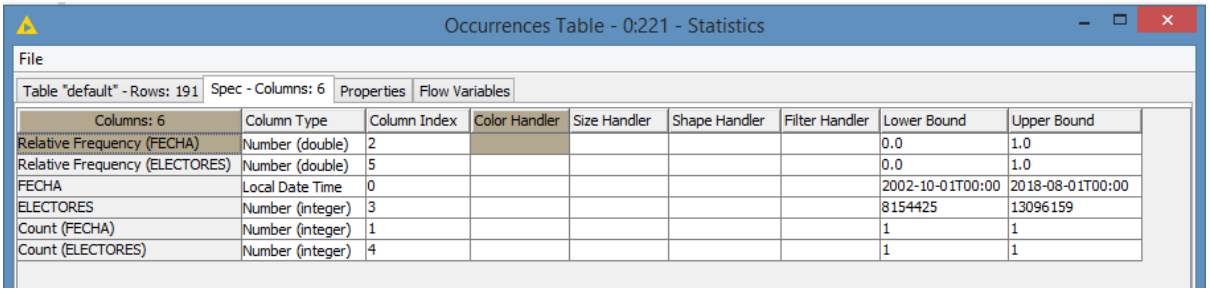

Figura 22. Información de los datos en el componente Statistics

A continuación, en la Figura 23 se visualiza el componente **Data Explorer** que ofrece una gama de opciones para mostrar las propiedades de los datos de entrada en una vista interactiva. Este componente es similar a Statistics la diferencia es la forma de presentación de la información al usuario final.

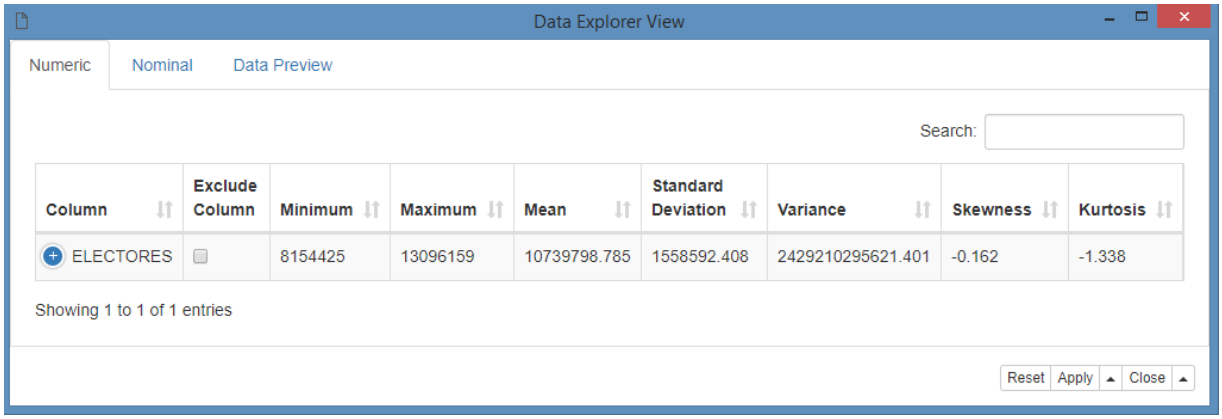

Figura 23. Utilización del componente Data Explorer

FASE III: Preparación de los datos

En esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, éstas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos. La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (Chapman et al., 2000).

Las actividades realizadas en esta fase son:

a) **Seleccionar los datos.** En esta tarea se seleccionó un subconjunto de los datos adquiridos desde la base de datos Oracle 12c, la información seleccionada fue desde enero de 2002 hasta agosto de 2018 segmentada mensualmente, es decir en total se extrajo 2 campos: Fecha (Mes/Año) y cantidad de electores, la cantidad seleccionada en total fue de 190 registros.

b) **Limpiar los datos.** Esta tarea complementa a la anterior y es una de las que más tiempo y esfuerzo consume debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos y posteriormente pasarlo a la fase de modelación, sin embargo, debido a que la información es enviada mediante web services por el Registro Civil y receptada por la Dirección Nacional de Sistemas e Informática Electoral del CNE, dentro de estos pasos ya se realiza una limpieza de datos antes de almacenar la información en la base de datos Oracle 12c, a pesar de aquello se verifico que no existan valores nulos tanto en el campo Fecha(Mes/Año) y cantidad de electores.

c) **Formateo de los datos.** En esta tarea principalmente se realizó las transformaciones sintácticas de los datos sin modificar su significado de tal forma que se permita y se facilite utilizar alguna técnica de minería de datos en concreto, por tal motivo se realizó el cambio del tipo de dato a *Date* en la columna Fecha (Mes/Año) debido a que estaba llegando desde la base de datos como string y la ordenación en forma ascendente de dicho campo.

A continuación, se observa en la Figura 24, la preparación de datos realizados.

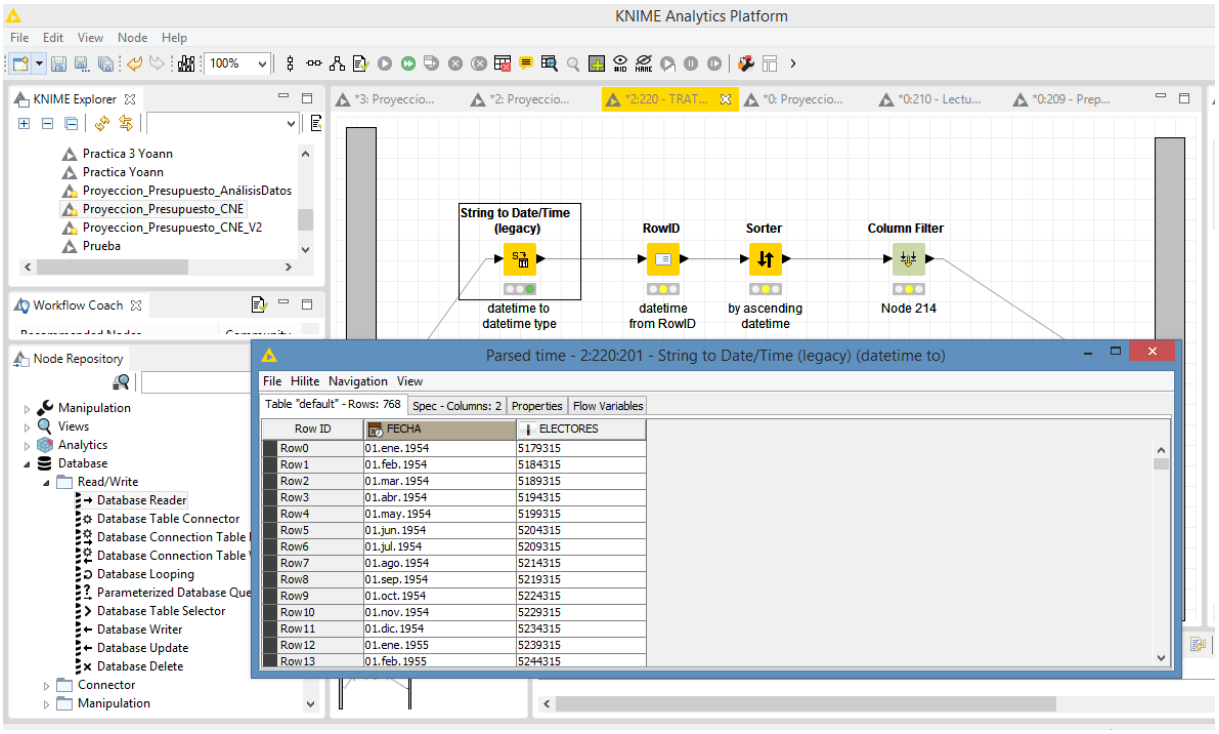

Figura 24. Agrupación mensual de información disponible de electores

En resumen, en la Figura 25 se sintetiza los procesos realizados en cada tarea de la fase de preparación de los datos.

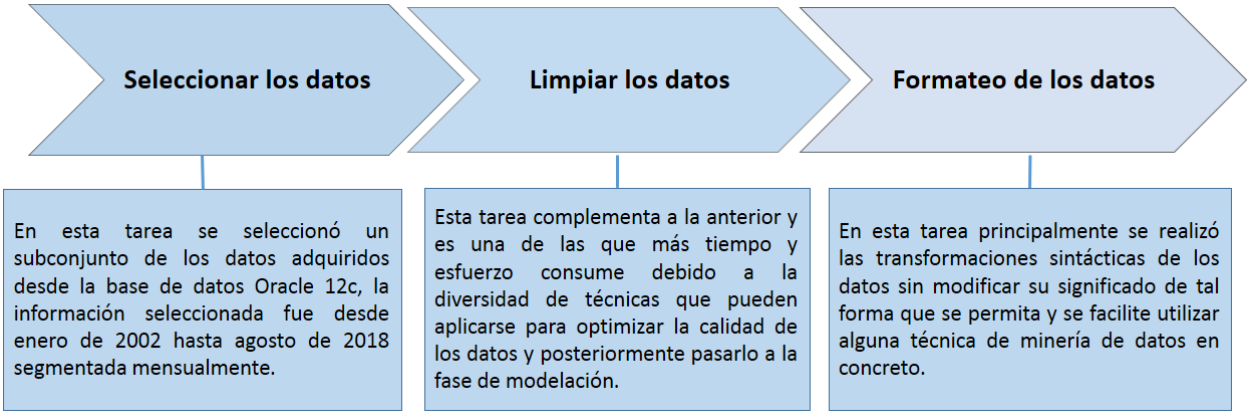

Figura 25. Resumen de las tareas realizadas en la fase de preparación de los datos.

FASE IV: Modelado

En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico.

Este modelado debe determinar un método de evaluación de los modelos que permita establecer el grado de adecuación de cada uno de ellos (Chapman et al., 2000).

Las actividades realizadas en esta fase son:

a) **Seleccionar la técnica de modelado.** En el capítulo 2, apartado 2.1.1 se realizó el análisis y selección de la técnica de minería de datos a utilizar, para esta selección se consideró el objetivo principal del proyecto en el Consejo Nacional Electoral del Ecuador (CNE) que es la predicción del presupuesto electoral, por tal motivo debido a que nuestro problema está relacionado con predicción se seleccionó la técnica de minería de datos predictiva.

   De la misma manera se realizó el análisis y selección del algoritmo analítico a utilizar dentro del proyecto de investigación, eligiendo el modelo de regresión por ser el mejor evaluado y utilizado en proyectos de analítica, el detalle de la selección se evidencia en el capítulo 2, apartado 2.2.1.

   Finalmente, la técnica y algoritmo a utilizar será implementado mediante la ayuda de una herramienta analítica llamada Knime Analytics Platform, la misma que contiene componentes analíticos incluidos en el software, el análisis detallado se encuentra en el capítulo 2, apartado 2.3.1.

b) **Generar el plan de prueba.** En esta tarea se genera un procedimiento destinado a probar la calidad y validez del modelo construido, por tal motivo con la información de la base de datos Oracle 12c en donde se encuentra almacenado la información mensual de electores

se realizó una separación de los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba, en la Figura 26, se evidencia que se aplicó un particionamiento del 90% para el aprendizaje del modelo y el 10% restante para la evaluación, esto con el fin de tener el nivel de calidad de la predicción.

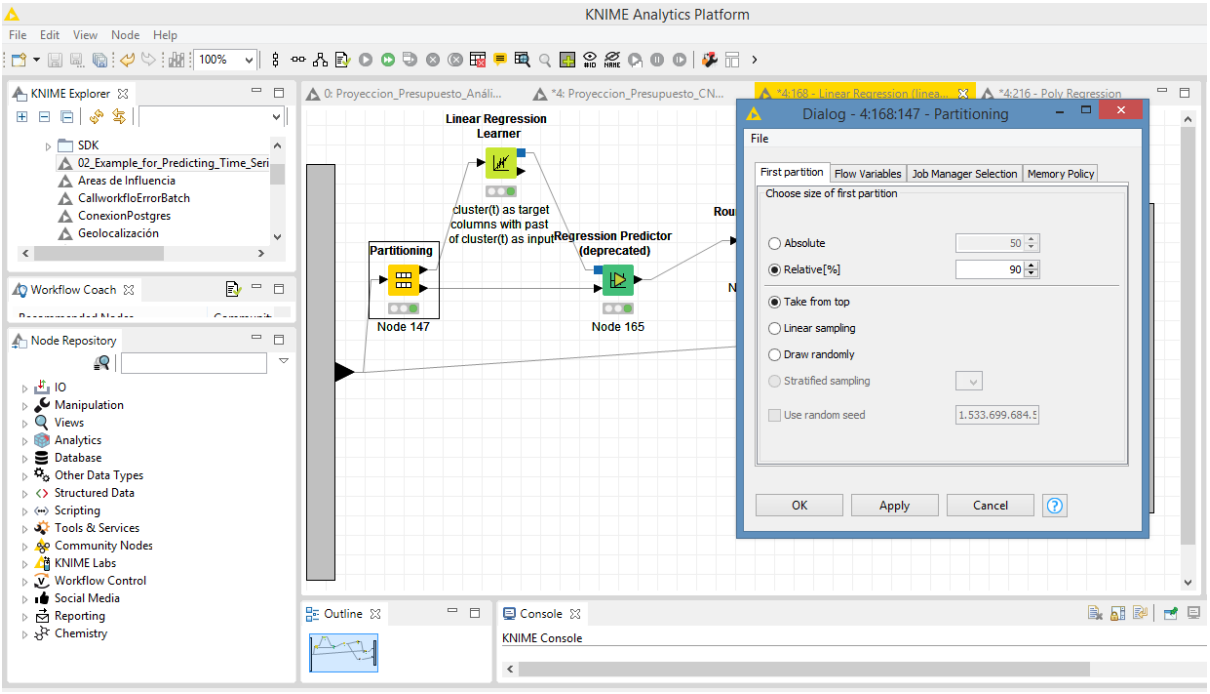

Figura 26. Generación del plan de prueba en el modelo

a) Construir el modelo. Para la construcción del modelo de proyección del presupuesto electoral se utilizó la técnica predictiva, el algoritmo de regresión aplicado dentro de la herramienta analítica Knime Analytics Platform, las mismas que fueron seleccionadas en el capítulo 2. En el presente trabajo de investigación se aplicó la técnica de análisis de regresión y se generaron dos modelos, uno con el algoritmo de regresión lineal y otro con el algoritmo de regresión Polinomial. En la Figura 27, se puede visualizar las

configuraciones que se aplicó a las variables correspondientes a fin de poder evaluar y seleccionar el rendimiento de cada uno de ellos.

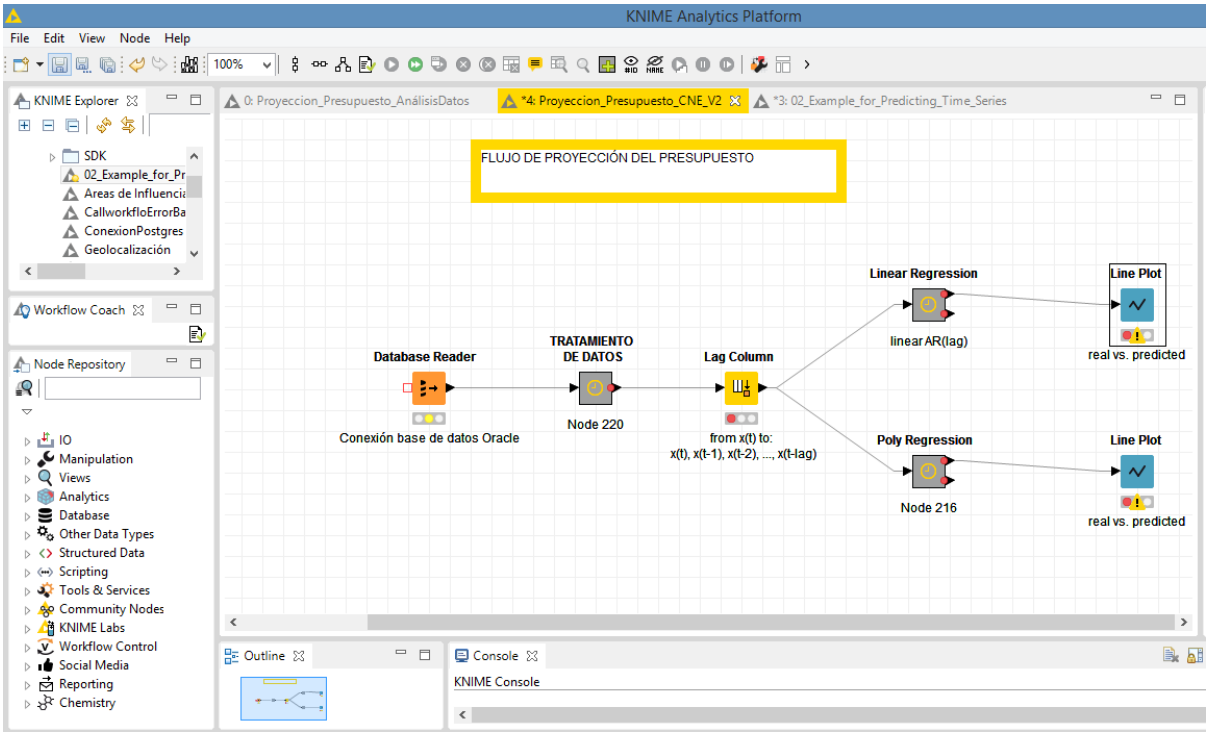

Figura 27. Construcción del modelo de regresión lineal y polinomial

b) Evaluar el modelo. Para esta última tarea de la fase de modelado se incluye un componente de evaluación disponible en la herramienta analítica Knime llamado “Numeric Scorer”, este componente permite la visualización del error del modelo, la cual ayudará a calcular el porcentaje de precisión y tomar la decisión de aceptar o rechazar el mismo, en la Figura 28 se expone la configuración del componente “Numeric Scorer” en donde se especifica el valor real extraído desde la base de datos Oracle 12c llamado “electores” y el valor que el modelo predijo llamado “Prediction (ELECTORES)”, con los valores mencionados la herramienta exporta el error del modelo analítico.

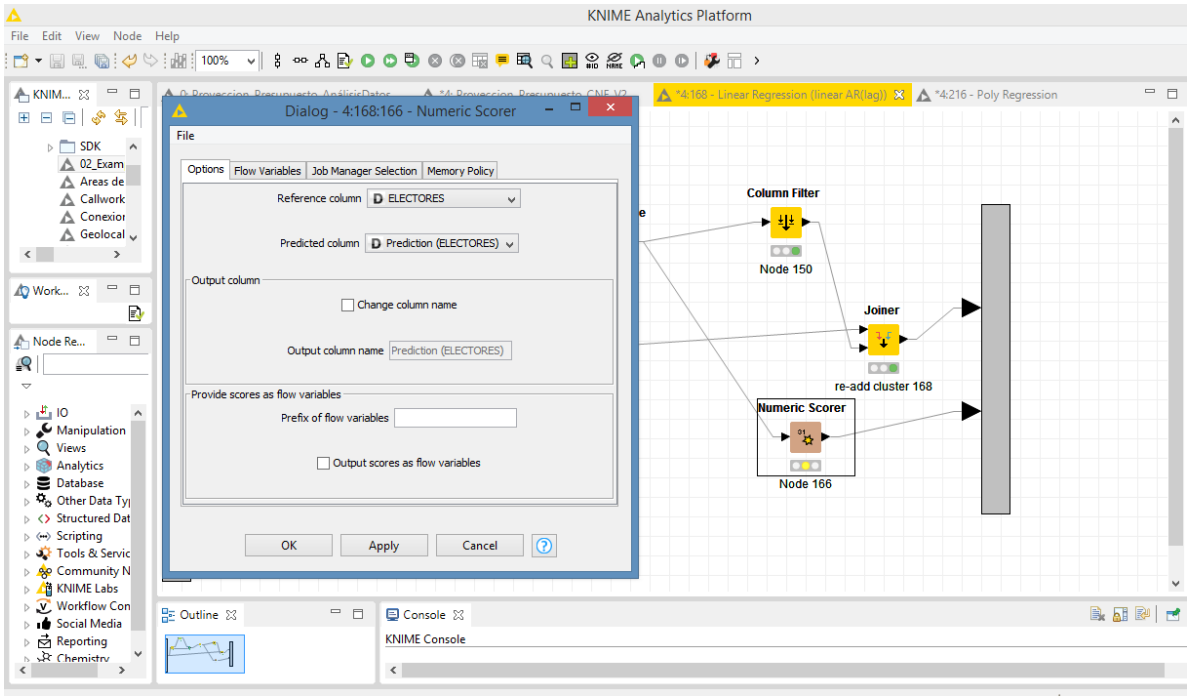

Figura 28. Módulo Scorer para la evaluación del modelo

En resumen, en la Figura 29 se sintetiza los procesos realizados en cada tarea de la fase de modelado.

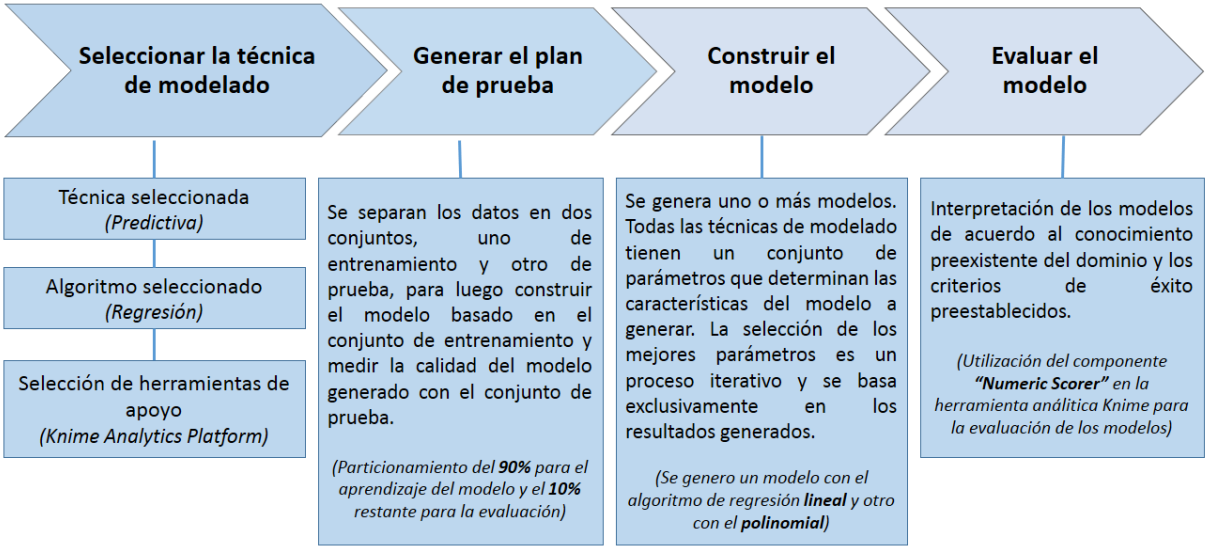

Figura 29. Resumen de las tareas realizadas en la fase de modelado

FASE V: Evaluación

En esta fase se evalúa el modelo, teniendo en consideración el cumplimiento de los criterios de éxito del problema, cabe mencionar que en el capítulo 1 nos planteamos resolver la siguiente hipótesis:

"Las técnicas de patrones de comportamiento permiten mejorar las predicciones en los indicadores presupuestarios del material electoral que maneja la Dirección Nacional de Logística."

La pregunta expuesta tendrá respuesta una vez que evaluemos el modelo y mejoremos el promedio del 25% de precisión que actualmente presenta la Dirección Nacional de Logística realizando este proceso mediante cálculos en Excel basado en el conocimiento del personal de esta dirección.

Debe considerarse además que la fiabilidad formulada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se pueda haber cometido algún error (Chapman et al., 2000).

Las actividades realizadas en esta fase son:

a) **Evaluar los resultados.** En los pasos anteriores de evaluación se trataron factores tales como la exactitud y generalidad del modelo. Esta tarea involucra la evaluación del modelo en relación a los objetivos del proyecto planteado que es mejorar el 25% de precisión que actualmente presenta la Dirección Nacional de Logística, por tal motivo se analizó el error de los modelos generados, el de regresión lineal y regresión polinomial, de tal manera que se pueda verificar cuál de ellos se adapta mejor a los objetivos propuestos al inicio del proyecto.

En la Figura 30, se visualiza el porcentaje de error absoluto del modelo de regresión lineal, determinado en un **10.42%.**

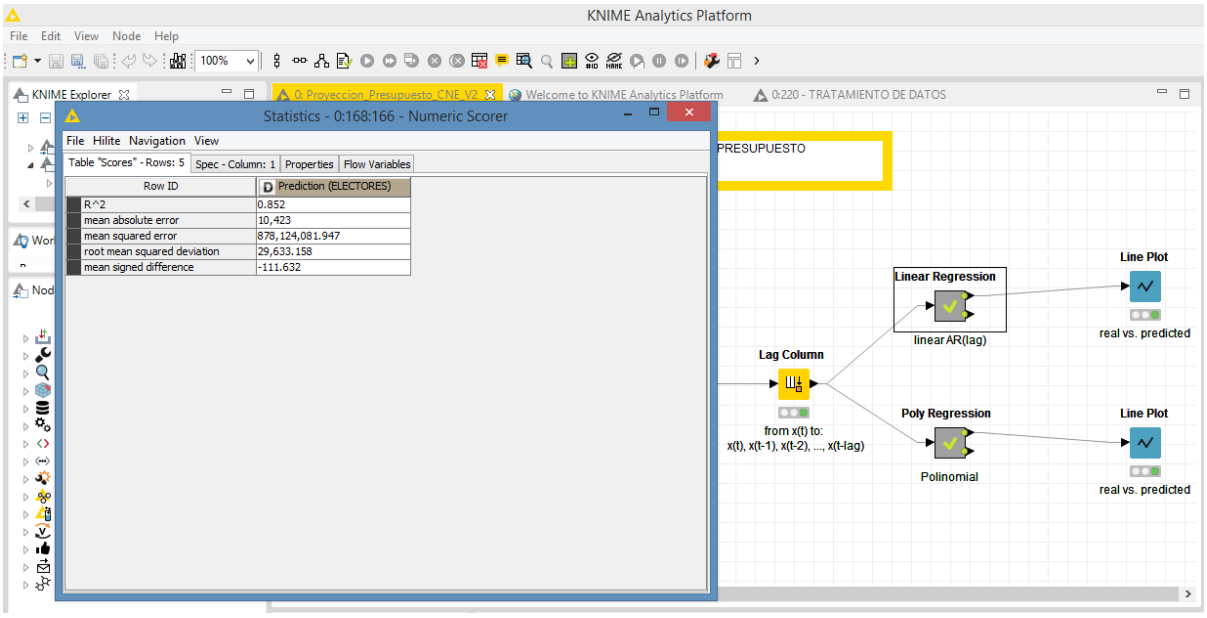

Figura 30. Evaluación del modelo de Regresión Lineal

De la misma manera se analiza el error absoluto del modelo de regresión polinomial, determinado en un **12.39%,** el cual se visualiza en la Figura 31.

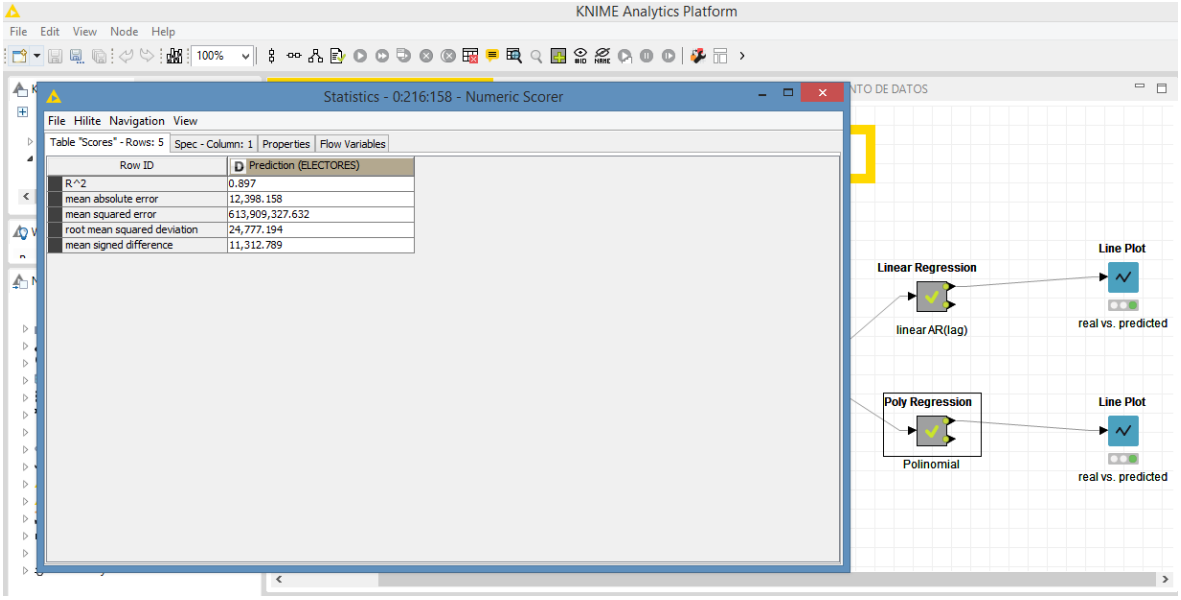

Figura 31. Evaluación del modelo de Regresión Polinomial

En la Tabla 7, se visualiza un comparativo del error absoluto obtenidos en los modelos de regresión lineal y regresión polinomial.

Tabla 7

Comparación de los algoritmos analíticos

Comparación de algoritmos		
Algoritmo	**Error Absoluto**	**Precisión (%)**
Regresión Lineal	10,42	**89,58**
Regresión Polinomial	12,39	**87,61**

En conclusión, el error absoluto del modelo de regresión lineal fue de **10.42%**, y el de regresión polinomial de **12.39%**, de acuerdo a los valores analizados podemos determinar que el modelo con menor porcentaje de error absoluto es el de regresión lineal, obteniendo un nivel de acierto del modelo de **89.58%**, visualizado en la Figura 30. Si realizamos la comparación con el acierto del 25% que actualmente lo realiza el área de logística del CNE, este valor mencionado por la Ing. Laura Ayala en la minuta de reunión del inicio del proyecto, expuesta en la Tabla 9 del Anexo 1, podemos concluir que el modelo generado de regresión lineal cumple el objetivo de mejorar la proyección del presupuesto electoral.

b) Revisar el proceso. Se refiere a calificar al proceso entero de minería de datos con el objeto de identificar elementos que pudieran ser mejorados, esta tarea de la fase de evaluación no se realizó debido a que el proceso cumple los objetivos propuestos al inicio del proyecto de investigación que es mejorar la proyección del presupuesto electoral definido actualmente en un promedio del 25% de acierto según lo mencionada la Ing. Laura Ayala, encargada de la Dirección de Logística del CNE (Ver Tabla 9-Anexo1. Minuta de reunión de inicio del proyecto).

c) **Determinar los próximos pasos.** Debido a que s**e** ha determinado que las fases anteriores han generado resultados satisfactorios se procede al análisis de la siguiente fase de la metodología.

En resumen, en la Figura 32 se sintetiza los procesos realizados en cada tarea de la fase de evaluación del modelo.

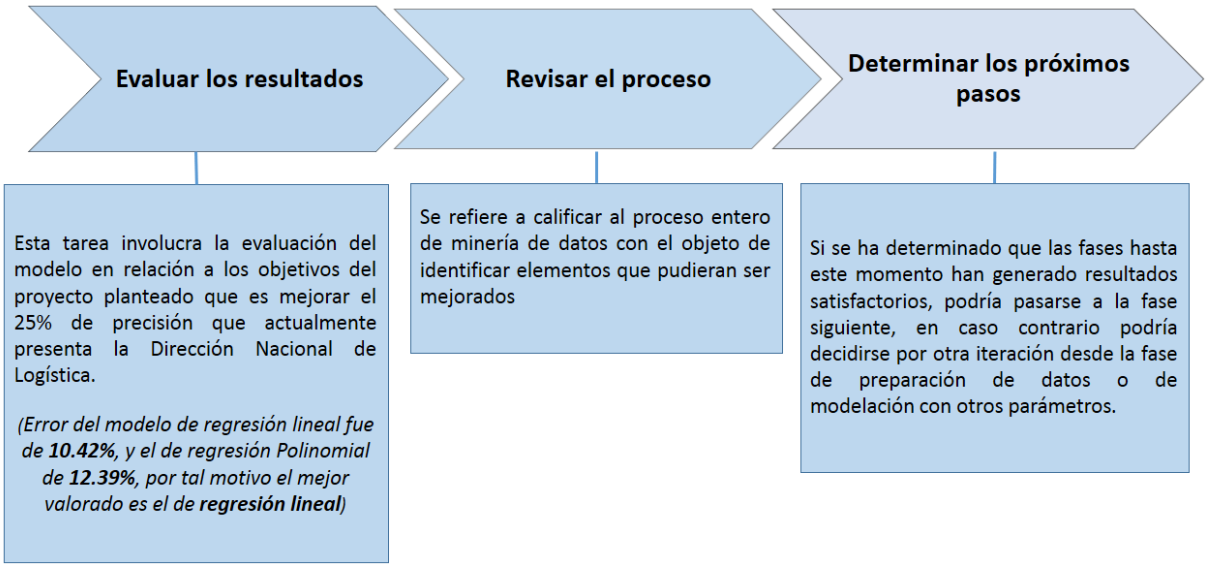

Figura 32. Resumen de las tareas realizadas en la fase de evaluación

FASE VI: Implantación

En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados. Generalmente un proyecto de minería de datos no concluye en la implantación del modelo, ya que se deben documentar y presentar los resultados de manera comprensible para el usuario con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de implantación se

debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados (Chapman et al., 2000)

Las actividades realizadas en esta fase son:

a) **Plan de implantación.** Para llevar a cabo este proyecto en la Institución es necesario que se cuente con acceso directo a la base de datos de Oracle 12c, con la finalidad de acceder a la información necesaria, se debe crear una base de respaldo para no tener conflicto en la base transaccional. Luego se seguirá con la fase de análisis de datos hasta la evaluación de los resultados, que se ha venido desarrollando durante la investigación. Otro punto corresponde a tener un servidor solo dedicado al análisis de datos y que tenga instalado la herramienta analítica Knime para ejecutar los modelos construidos durante la investigación. Este plan se lo debe realizar 1 año antes de cada proceso electoral para obtener la predicción del presupuesto electoral de la Dirección de Nacional de Logística, debido aquello se ha generado un informe del proceso de implantación y puesta en marcha del proyecto de predicción del presupuesto electoral, el proceso a seguir se describe en el Anexo 2 del presente trabajo de investigación.
b) **Plan de monitoreo y mantenimiento.** Luego de que se haya implementado el proyecto, el Consejo Nacional Electoral del Ecuador realizará un monitoreo por el lapso de 3 meses para la verificación del funcionamiento del proyecto, siempre que no se cambien los objetivos de la minería de datos establecidos en esta investigación.

En conclusión, en esta fase se dejó la documentación correspondiente para que la institución pueda realizar la implantación en producción del proyecto analítico, luego que hayan realizado el monitoreo para la verificación del funcionamiento del proyecto.

En resumen, la metodología CRISP-DM se sintetiza en 6 fases descritas con sus tareas correspondientes en la Figura 33.

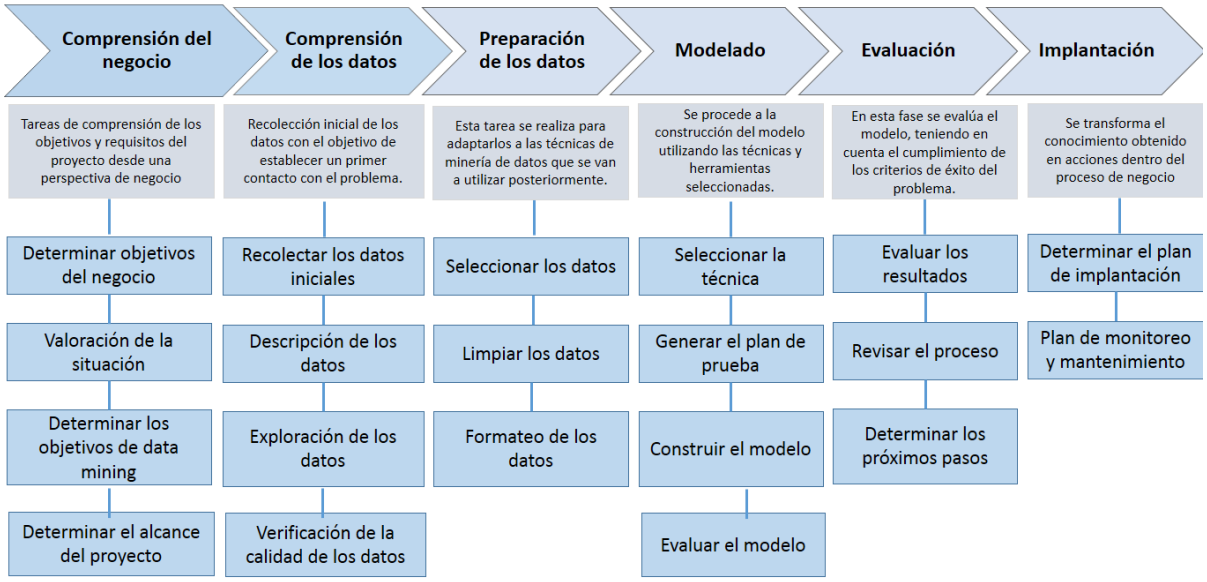

Figura 33. Resumen de las fases de la metodología CRISP-DM

CAPITULO IV

VALIDACIÓN DEL MODELO ANÁLITICO

En esta sección se realiza una validación del modelo más detallado, debido a que en el anterior capítulo se realizó una evaluación dentro de la metodología aplicada.

Para la validación del modelo se utilizó la herramienta analítica Knime Analytics Platform debido a que incluye componentes de certificación de la veracidad de modelos predictivos.

4.1 Validación del modelo de Regresión Lineal

Se ha realizado una validación a través de la herramienta analítica Knime Analytics Platform, la misma que arroja un error absoluto del 10,42%, como se visualiza en la Figura 34, se ha tomado como referencia este error debido a que es el más utilizado para medir la precisión en modelos predictivos (Velásquez H., Dyner R., & Souza., , 2006).

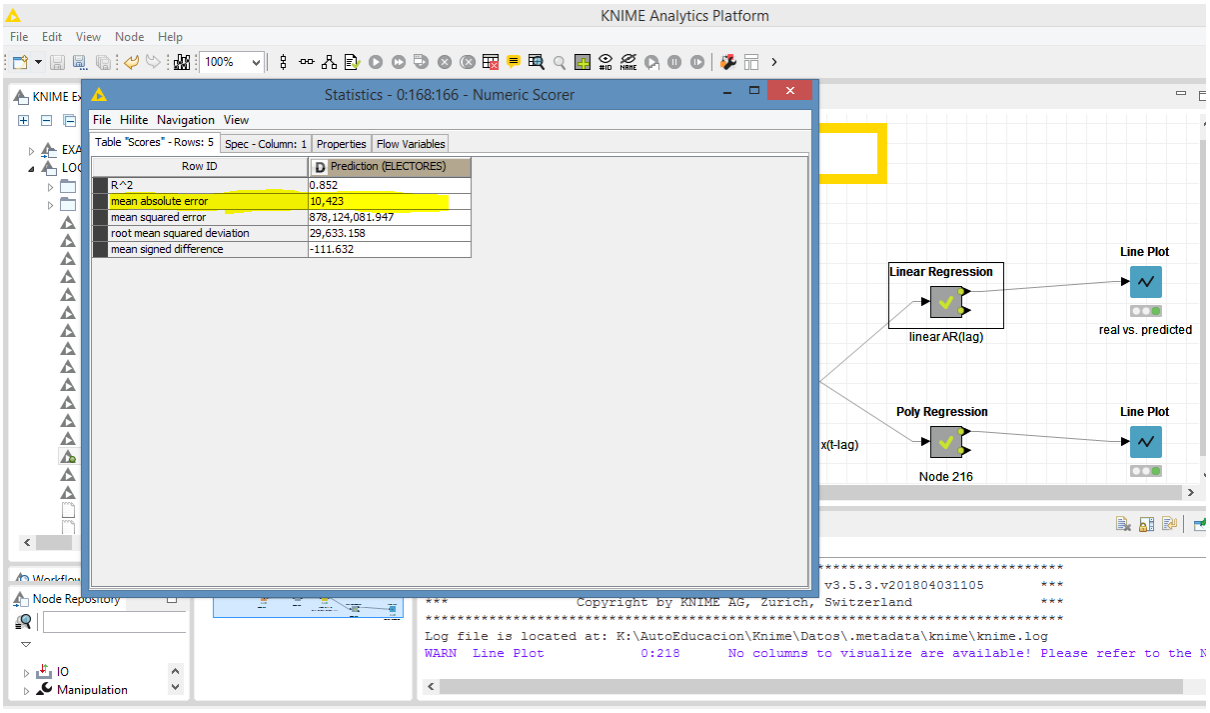

Figura 34. Evaluación del modelo de Regresión Lineal

Si realizamos el cálculo de precisión que se obtiene aplicando este modelo podemos verificar que presentamos un **89,58%** de acierto como se evidencia en los siguientes cálculos.

$$\textbf{\textit{Precisión del Pronóstico}}\ (\%) = \left(100\% - \left(\frac{abs(venta\ real\ - forecast)}{venta\ real} * 100\%\right)\right)$$

$$\textbf{\textit{Precisión del Pronóstico}}\ (\%) = (100\% - \text{Error Absoluto})$$

$$\textbf{\textit{Precisión del Pronóstico}}\ (\%) = (100\% - 10{,}42\%) = \mathbf{89{,}58}\ \%$$

4.2 Validación del modelo de Regresión Polinomial

De la misma manera se utilizó el cálculo del error absoluto (Velásquez H. et al., 2006) aplicado en el algoritmo de Regresión Polinomial obteniendo un valor de **12,39%** en el error, como se evidencia en la Figura 35.

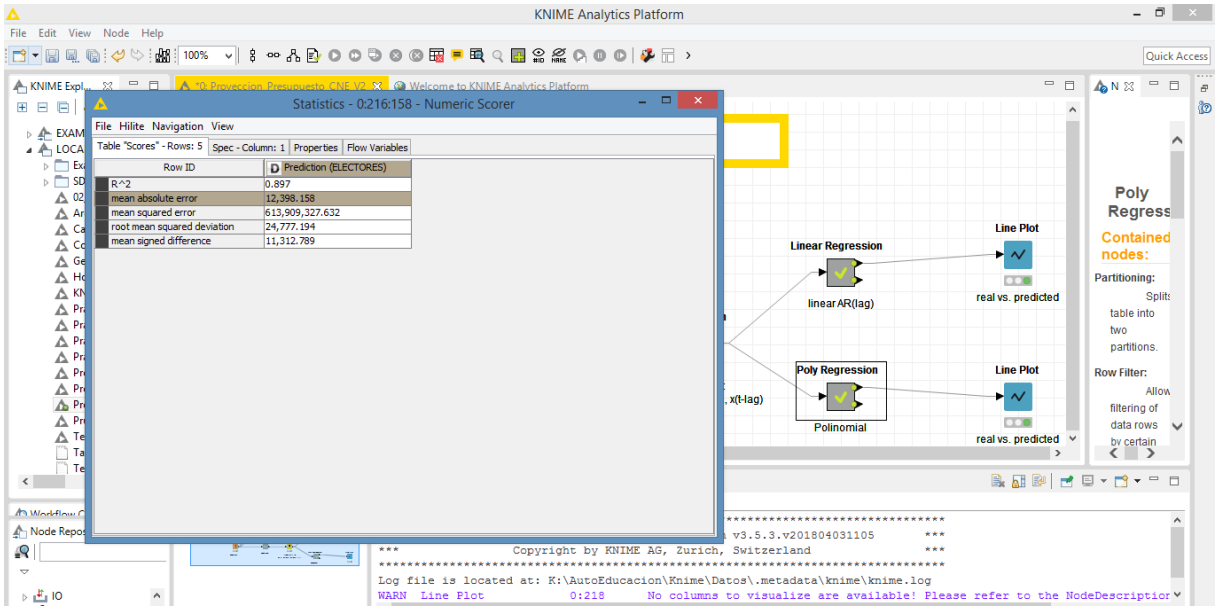

Figura 35. Evaluación del modelo de Regresión Polinomial

Al realizar el cálculo de precisión del algoritmo se obtuvo un valor de **87,61%** de acierto, como se evidencia en el detalle del siguiente proceso de cálculo:

$$\textbf{\textit{Precisión del Pronóstico}}\ (\%) = \left(100\% - \left(\frac{abs(venta\ real\ - forecast)}{venta\ real} * 100\%\right)\right)$$

$$\textbf{\textit{Precisión del Pronóstico}}\ (\%) = (100\% - \text{Error Absoluto})$$

$$\textbf{\textit{Precisión del Pronóstico}}\ (\%) = (100\% - 12{,}39\%) = \mathbf{87{,}61}\ \%$$

4.3 Comparativo de la validación de los algoritmos analíticos

Como una forma de poder comparar los dos algoritmos analíticos se realiza la tabla comparativa con sus valores absolutos y el porcentaje de precisión obtenido en cada uno de ellos, en donde se puede visualizar que el algoritmo con mayor valor de precisión es el de regresión lineal determinada en **89,58%**, descrita en la Tabla 8.

Tabla 8

Validación de los algoritmos analíticos

Comparación de algoritmos		
Algoritmo	**Error Absoluto**	**Precisión (%)**
Regresión Lineal	10,42	**89, 58**
Regresión Polinomial	12,39	**87, 61**

De acuerdo al benchmark realizado de la precisión de los dos algoritmos: Regresión Lineal y Regresión Polinomial, se puede concluir finalmente que el algoritmo de Regresión Lineal es el mejor valorado en el presente trabajo de investigación y el cuál arroja los mejores resultados.

CONCLUSIONES

- Se concluyó que las técnicas de patrones de comportamiento permiten mejorar las predicciones en los indicadores presupuestarios del material electoral que maneja la Dirección Nacional de Logística en el Consejo Nacional Electoral CNE. Se validó el modelo predictivo del presupuesto y se obtuvo un promedio de precisión del 89,50%. En base a estos datos se puede mencionar que la hipótesis planteada es verdadera, debido a que mejora notablemente la precisión que maneja el Consejo Nacional Electoral del Ecuador que es del 25% sin aplicar la solución propuesta.
- El proyecto analítico se realizó con la metodología CRISP-DM, con Knime Analytics Platform y Regresión Lineal que es la herramienta y algoritmo estadístico mejor valorado de acuerdo a los criterios de evaluación descritos en el presente trabajo de investigación, y con el cual el Consejo Nacional Electoral del Ecuador obtuvo una implementación rápida, fácil y con resultados de acierto eficientes para este proceso.
- Se realizó la proyección del presupuesto electoral en el CNE obteniendo los resultados esperados de mejorar los procesos manuales que actualmente lleva la institución, utilizando herramientas y algoritmos líderes en el mercado y los mejores valorados en el presente trabajo de investigación, con el proyecto de minería de datos mencionado se efectuó una inducción de analítica avanzada dentro de la institución, un área aún no explorada y que poco a poco tomara fuerza.

RECOMENDACIONES

- Se recomienda al Consejo Nacional Electoral la implantación de una base de datos analítica, debido a que con el crecimiento de los datos la conexión a una base de datos transaccional no es lo recomendable debido a que perjudica el rendimiento del mismo.
- El modelo analítico propuesto en el presente trabajo de investigación debería ser utilizado e implementado en producción debido a que automatizará y optimizará los procesos de toma de decisiones.

BIBLIOGRAFÍA

Arredondo Vidal. (2009). *Introducción a los Algoritmos Genéticos.* Obtenido de http://profesores.elo.utfsm.cl/~tarredondo/info/soft-comp/Introduccion%20a%20los%20Alg

Blacke & Jacobson. (2016). *Arquitectura de negocio de FORD.* Obtenido de https://pages.alteryx.com/Ford-enabling-transformational-change-TY.html?aliId=188728387

Bulkley. (2018). Western Union uses Alteryx to Stay Compliant and Improve Customer Insight.

Cano. (2007). *Business Intelligence, Competir con Información*.

Chapman et al. (2000). *CRISP-DM 1.0.* Obtenido de https://www.the-modeling-agency.com/crisp-dm.pdf

Díaz. (2016). *Metodología CRISP-DM - Parte 1.*

Forrester. (2018). *About Forrester.* Obtenido de https://www.forrester.com/marketing/policies/forrester-wave-methodology.html

Gartner. (2018). Magic Quadrant for Data Science and Machine-Learning Platforms. 442.

Gregory Piatetsky. (2016). *Algoritmos principales y métodos utilizados por los científicos de datos.* Obtenido de https://www.kdnuggets.com/2016/09/poll-algorithms

Gregory Piatetsky. (2017). *Forrester vs Gartner on Data Science Platforms and Machine Learning Solutions.* Obtenido de https://www.kdnuggets.com/2017/04/forres

halweb. (2013). *Series Temporales.* Obtenido de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema7.pdf

hfrancob. (2011). Arquitectura De Business Intelligence - Parte 1.

IT Ahora. (2018). Analítica: El futuro de la región es un desarrollo personalizado.

IT Ahora. (2018). Seguros: La forma de hacer negocios está cambiando.

itahora. (2018). Analítica de Datos Avanzada y visualización generan oportunidades de negocio en AYASA.

KDnuggets. (2014). *CRISP-DM, metodología para análisis, extracción de datos o proyectos de ciencia de datos.* Obtenido de https://www.kdnuggets.c

Kdnuggets. (2018). *Gregory Piatetsky-Shapiro*. Obtenido de https://www.kdnuggets.com/gps.html

KNIME. (2018). *KNIME Analytics Platform.* Obtenido de https://www.knime.com/knime-software/knime-analytics-platform

Microsoft. (2018). *Algoritmos de minería de datos (Analysis Services: minería de datos).* Obtenido de https://docs.microsoft.com/es-es/sql/analysis-servic

Moine et al. (2011). *Estudio comparativo de metodologías para minería de datos.* Obtenido de https://digital.cic.gba.gob.ar/bitstream/handle/11746/3525/Estudio

Piatetsky. (2018). *Gainers and Losers in Gartner 2018 Magic Quadrant for Data Science and Machine Learning Platforms*. Obtenido de https://www.kdnuggets.com/2018/02/gartne

Piatetsky-Shapiro. (2018). *Acerca de KDnuggets, Analytics, Big Data, Data Mining y Data Science leader.* Obtenido de https://www.kdnuggets.com/about/index.html

Rodriguez, Álvarez, Mesa, & González, . (2003). Metodologías para la realización de Proyectos de Data Mining.

Santín Gonzalez & Peréz Lopéz. (2008). *Minería de datos. Técnicas y herramientas.*

Scopus. (2018). *Scopus - Analyze search results.* Obtenido de https://www-scopus-com.bibliotecavirtual.udla.edu.ec/term/analyzer.uri?sid=074c8d3dee300be7319578158ec39884&orig

SIAG Consulting. (2016). ¿Qué es exactamente el famoso cuadrante mágico de Gartner?

UCM. (2018). *Métodos Bayesianos.* Obtenido de http://www.mat.ucm.es/~villegas/info/bayesianos/

Underwood. (2017). How adidas Transformed Analytics Platforms for Digital Scale.

Velásquez H., Dyner R., & Souza., . (2006). TENDENCIES IN THE PREDICTION AND ESTIMATION OF THE CONFIDENCE INTERVALS USING MO.

WebMining Consultores. (2011). *KDD Proceso de Extracción de conocimiento.* Obtenido de http://www.webmining.cl/2011/01/proceso-de-extraccion-de-con

Zaldivar et al. (2011). *Comparativa entre los Métodos de Regresión Lineal y Minería de Datos para la Identificación de Variables Asociadas al Éxito Académico* .