ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN Y TRANSFERENCIA DE TECNOLOGÍA DIRECCIÓN DE POSTGRADOS

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

TEMA: "ANÁLISIS DE TENDENCIAS Y DESCUBRIMIENTO DE PATRONES DE COMPORTAMIENTO DE DEMANDAS JUDICIALES PARA EL CONSEJO DE LA JUDICATURA UTILIZANDO ALGORITMOS Y TÉCNICAS DE MINERÍA DE DATOS"

AUTOR: ING. MARTEL SOCOLA, WILFREDO IVAN

DIRECTOR: ING. ECHEVERRÍA BRÍONES, PEDRO FABRICIO MSG.

SANGOLQUÍ - ECUADOR

2017

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE MAESTRÍA EN SISTEMAS DE GESTIÓN DE LA INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

CERTIFICACION

Certifico que el trabajo de titulación, **"ANÁLISIS DE TENDENCIAS Y DESCUBRIMIENTO DE PATRONES DE COMPORTAMIENTO DE DEMANDAS JUDICIALES PARA EL CONSEJO DE LA JUDICATURA UTILIZANDO ALGORITMOS Y TÉCNICAS DE MINERÍA DE DATOS"** realizado por el señor **WILFREDO IVAN MARTEL SOCOLA**, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar al señor **WILFREDO IVAN MARTEL SOCOLA** para que lo sustente públicamente.

Sangolquí, Febrero del 2017

MSG. PEDRO FABRICIO ECHEVERRÍA BRÍONES
DIRECTOR

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE MAESTRÍA EN SISTEMAS DE GESTIÓN DE LA INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORÍA DE RESPONSABILIDAD

Yo, **WILFREDO IVAN MARTEL SOCOLA**, con cédula de identidad **Nº 070520659-7** declaro que este trabajo de titulación **"ANÁLISIS DE TENDENCIAS Y DESCUBRIMIENTO DE PATRONES DE COMPORTAMIENTO DE DEMANDAS JUDICIALES PARA EL CONSEJO DE LA JUDICATURA UTILIZANDO ALGORITMOS Y TÉCNICAS DE MINERÍA DE DATOS"** ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 01 de febrero del 2017

WILFREDO IVAN MARTEL SOCOLA
C.C. 0705206597

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE MAESTRÍA EN SISTEMAS DE GESTIÓN DE LA INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORIZACIÓN

Yo, **WILFREDO IVAN MARTEL SOCOLA**, No autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca virtual de la institución el presente trabajo de titulación "**ANÁLISIS DE TENDENCIAS Y DESCUBRIMIENTO DE PATRONES DE COMPORTAMIENTO DE DEMANDAS JUDICIALES PARA EL CONSEJO DE LA JUDICATURA UTILIZANDO ALGORITMOS Y TÉCNICAS DE MINERÍA DE DATOS**" cuyo contenido, ideas y criterios de mi autoría y responsabilidad.

Sangolquí, 01 de febrero del 2017

WILFREDO IVAN MARTEL SOCOLA
C.C. 0705206597

DEDICATORÍA

A Dios, porque, quien lo lleva a su lado nunca le faltará nada.

AGRADECIMIENTOS

En primer lugar, quisiera a gradecer a mis padres y a mis hermanas, por enseñarme a ser siempre persistente para conseguir mis sueños; me enseñaron que luchar siempre por nuestros ideales es el sentido de la vida.

En segundo lugar, quiero agradecer infinitamente a mi tutor Ing. Fabricio Echeverría MSG. por su confianza y su constante revisión a mi tesis de maestría.

Gracias a todos ustedes.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La presente investigación se llevó a cabo en la Función Judicial del Ecuador en el año 2016, con la finalidad de ayudar a la institución en brindar una herramienta a los ecuatorianos y ecuatorianas del país, ya que por el desconocimiento de leyes que existe por parte de las personas con respecto a la duración de los procesos judiciales, otros se aprovechan de esta particularidad para lucrarse y otros para hacer que las personas desistan de las demandas que recaen sobre ellos, por este motivo se desarrolló un modelo predictivo que analizan la duración de finalización de las demandas y las tendencias de las mismas para estimar la duración promedio de una causa procesal y ofrecer ese resultado valioso a los usuarios finales que les pueda servir como referencia para sus decisiones judiciales. Además, para llevar a cabo la investigación y no desfasarse del proyecto se aplicó la metodología CRISP-DM, en la cual se inició con el proceso de entender el negocio para posteriormente proceder con la exploración de los datos y obtener una radiografía de ellos, después de eso de procedió a verificar su calidad antes de empezar el análisis. Una vez que se cumplieron esos pasos, se procedió con el desarrollo de los modelos predictivos para cada uno de los objetivos definidos, y, se pasó a evaluarlos para seleccionar el mejor. Finalmente se emite un resultado de la evaluación en conjunto con las recomendaciones y conclusiones que se deberían seguir para el correcto funcionamiento y aplicación del mismo.

Palabras Clave:

- **MINERÍA DE DATOS**
- **MODELO DE MAQUINA DE SOPORTE**
- **MODELO LINEAL**
- **MODELOS PREDICTIVOS A LA FUNCIÓN JUDICIAL.**

ABSTRACT

The present research was carried out in the Función Judicial of Ecuador in the year 2016, with the purpose of helping the institution to provide a tool to the Ecuadorians, because of the lack of knowledge that exists on the part of the people with regard to the length of judicial proceedings, others take advantage of this particularity to make a profit and others to make people desist from the demands that fall on them, for this reason was developed a predictive model that analyze the duration of completion of the demands and tendencies of the same to estimate the average length of a procedural cause and to offer that valuable result to the end users that can serve as reference for their judicial decisions. In addition, the CRISP-DM methodology was applied in order to carry out the research and not to be out of step with the project, which began with the process of understanding the business and then proceeding with the exploration of the data and obtaining an X-ray of them. After that, it's proceeded to verify data quality before beginning the analysis. Once these steps were fulfilled, we proceeded with the development of predictive models for each of the defined objectives, and, then, models are evaluated to select the best one. Finally, a result of the evaluation is issued together with the recommendations and conclusions that should be followed for the correct functioning and application of the models.

Keywords:

- **DATA MINING**
- **VECTORIAL SUPPORT MACHINE MODEL**
- **LINEAR MODEL**
- **PREDICTIVE MODEL FOR FUNCION JUDICIAL**

CAPÍTULO I
ASPECTOS GENERALES

1.6.1 Introducción

En el mundo actual, es sorprendente la cantidad de información que se genera, ya no existe lugar para el almacenamiento físico de documentos porque lo más importante es tener todo informatizado y cuantificado en base de datos, pero más allá del almacenamiento, surge la necesidad de sacarle provecho a toda esa información almacenada para obtener conclusiones que ayuden a rentabilizar el negocio, obtener ventajas frente a los competidores y brindar un mejor servicio al consumidor.

Si bien es cierto, en la época actual mediante consultas simples sobre los datos se pueden obtener algunas respuestas, pero a conforme la complejidad aumenta en la base de datos y la enorme cantidad de registros es anormal, obtener los resultados se vuelven más difíciles de interpretar y analizar para la organización que desea utilizarlos para su propio beneficio. De esta necesidad nace la minería de datos que es la ciencia que estudia patrones y/o relaciones en grandes bases de datos y emplea técnicas de estadística, aprendizaje automático e inteligencia artificial para extraer la información y traducirla a unos resultados que puedan ser fácilmente interpretables por la persona o empresa que desea obtener provecho de los datos.

Así pues, para comenzar el proceso de minería de datos es importante partir de una base de datos que contenga toda la información que se quiera analizar y la misma debe estar correctamente estructurada para su uso posterior. La minería de datos trata de obtener conocimiento de toda la información posible de los registros de la base de datos, no se conforma sólo con la visualización de los datos como podría pasar con las consultas simples, si no que trata de obtener resultados en base a las relaciones, características de los datos que comparten entre si y de esta forma obtener ventajas y/o beneficios para el negocio.

Existen diversas técnicas y metodologías de minería de datos que se pueden utilizar y que pueden ser más o menos adecuadas para cada caso en concreto, pero en el presente proyecto se ha elegido la metodología CRISP-DM como guía para realizar la explotación de los datos contenidos en la base de datos de la Función Judicial, en la cual se analizará la información del inicio y finalización de las demandas judiciales y la tendencia de las mismas con el fin de tener pronósticos que ayuden a mejorar el servicio que se brinda a la ciudadanía.

Esta tesis está dividida en cuatro capítulos donde los dos primeros (la más teóricas) ponen en contexto al lector y darle una serie de conocimientos básicos acerca de la minería de datos y los distintos métodos que existen para facilitar el proceso de extracción de la información. Desde el capítulo tres en adelante es donde se aplica en la práctica las distintas etapas de la metodología escogida sobre los datos que se dispone para finalmente hacer una valoración lo más objetiva posible acerca de la viabilidad de del problema a resolver en este proyecto.

1.6.2 Antecedentes

Historiadores describen que la Función Judicial inicio su proceso de informatización en las cortes de justicia más importante del país cerca del año de 1999. Entre las ciudades donde se aplicó esta fase piloto son: Quito, Cuenca, Guayaquil, Riobamba entre otras. Durante este tiempo la justicia ha venido evolucionando a tal punto que en la actualidad es guía de referencia para países latinoamericanos tanto en términos de justicia como informáticos.

Tal como se había mencionado, la automatización de la Función Judicial se realiza a través del sistema “eSATJE” y éste está compuesto por una serie de módulos que se definen a continuación: sorteos y notificaciones electrónicas, cobranza y pago de pensiones, antecedentes penales y tránsito entre otros. Además, el aplicativo ya se aplica a todas las materias penales y no penales en todas las dependencias jurisdiccionales del país. (Función Judicial, 2016)

1.6.3 La institución

La Función Judicial como parte importante del estado está conformada por cortes y tribunales de justicia quienes velan porque se aplique la justicia de forma igualitaria a todos los rincones del país. (Revista Judicial de Derecho, 2016)

Visión. - hacer de la justicia ecuatoriana un sinónimo de calidad y confianza llena de valores que mire por los derechos de las personas tanto individual como colectivamente y, además, sea toma como referencia por su excelencia.

Misión. - Contribuir a la paz social del país mediante el servicio de una justicia efectiva, oportuna que cubra las necesidades jurídicas y sea accesible a todas las personas.

Principios Fundamentales de la Función Judicial:

- Honestidad
- Justicia independiente
- Imparcialidad
- Transparencia
- Compromiso y responsabilidad social

1.6.4 Reseña Histórica

En sus inicios, la sede principal del consejo de la judicatura se establecía en la ciudad de Quito donde desde allí ejercía la ley en todo el territorio ecuatoriano tal como lo definía la Constitución. En aquellos tiempos, lo que actualmente es la función judicial estaba administrada por el Presidente del Consejo Nacional en conjunto con sus vocales. Luego en marzo de 1988, se expide una ley orgánica que estaba conformada de la siguiente manera:

- El Presidente
- El Director Ejecutivo

- El Pleno y,
- Recursos Humanos junto con la Comisión Financiera.

Con esa nueva conformación en el Consejo de la Judicatura, se incrementan el número de vocales a nueve y se elige al doctor Xavier Arosemena Camacho como presidente la Institución. Con respecto a ocupar los puestos de los vocales se realiza un concurso de méritos y oposición en donde quedan los doctores Oscar León Guerrón y Homero Tinoco Matamoros como ganadores del concurso por obtener las mejores calificaciones.

Luego, se procedió a integrar la comisión Administrativa Financiera y la de Recursos Humanos. La parte de Recursos Humanos su presidente fue el doctor Hernán Proaño y sus vocales los doctores Xavier Arosemena Camacho, Benjamín Solórzano, Homero Tinoco Matamoros y Rosa Cotacachi Narváez. Y en la parte Administrativa el presidente fue Hernán Jaramillo Ordoñez y los vocales fueron los doctores: Jorge Vaca Peralta, Oscar León Guerrón y Ulpiano Salazar Ochoa. (Función Judicial, 2016)

1.6.5 Planteamiento del problema

Desde hace algunos años la justicia ha venido experimentando cambios positivos en los avances de los procesos judiciales. Y, asimismo, ha realizado grandes inversiones en la parte de tecnologías de información para su proceso de automatización. Actualmente la función judicial utiliza el sistema eSATJE para registrar el inicio y culminación de las causas judiciales y su éxito es innegable pero su preocupación por la automatización de ciertos procesos, que son esenciales para ella misma, ha conllevado al olvidado del primer factor clave que toda persona se pregunta ¿cuánto tiempo durará el juicio? .

La razón porque se realiza esta pregunta es por la problemática que existe hoy en día, los ciudadanos desconocen el tiempo promedio de duración de un proceso judicial, lo que se refleja mayoritariamente a desistir de los procesos legales produciendo que no sea práctico llevar un juicio ante la ley. Además, las personas

asocian el tiempo con dinero, lo que significa a mayor tiempo se prolongue el juicio mayor dinero será el que se tenga que gastar. Adicionalmente se tiene que agregar otros factores que intervienen en el proceso tal como los abogados, los cuales aprovechan el desconocimiento de los ciudadanos respecto a la duración de una causa judicial para prolongar el tiempo y lucrarse de ello. En resumen, mientras no exista una aplicación para estimar el tiempo promedio de una causa procesal y le indique al ciudadano la duración de su juicio, el pueblo tendrá la noción de que la justicia solo es para los que tienen dinero.

1.6.6 Formulación del problema

¿Cómo desarrollar un modelo predictivo basado en análisis y patrones de duración de las demandas judiciales que determine el tiempo de un determinado juicio en las materias no penales de la función judicial en el año 2015?

1.6.7 Preguntas de Investigación

¿Cómo verificar la existencia de la relación entre los propios registros de los procesos judiciales que determine un patrón de comportamiento en las demandas judiciales?

¿Cómo determinar la tendencia en las demandas judiciales realizadas en las materias no penales durante el año 2015?

¿Cómo determinar cuáles son los tipos de acción de mayor demanda en las materias no penales durante el año 2015?

1.6.8 Limitaciones y Supuestos

El proyecto se limita llegar hasta el desarrollo del modelo que cumpla con la predicción de tiempo promedio para las causas procesales de las materias no penales del año 2015 y a la vez propondrá las pautas a seguir para su implantación. No es posible implantar directamente el modelo en producción por cuestiones de limitación

de tiempo y la cantidad de aprobaciones que se debe tener para ponerlo servicio al público.

1.6.9 Objeto General

Desarrollar un modelo predictivo mediante el análisis de tendencias y patrones de duración de las demandas judiciales no penales para determinar el tiempo promedio de duración de un juicio en la ciudad de Quito-Ecuador en el año 2015.

1.6.10 Objetivos Específicos

- Verificar la existencia de la relación entre los propios registros de los procesos judiciales en la materia no penales del año 2015.
- Desarrollar un modelo predictivo de la duración promedio de un juicio a través del análisis de patrones de los procesos judiciales no penales del año 2015.
- Determinar las tendencias de los procesos judiciales no penales utilizando el historial de registros de los procesos judiciales del año 2015.
- Validar el modelo obtenido.
- Aplicar correctamente la metodología CRISP-DM para el desarrollo del proyecto de investigación.

1.6.11 Meta

Desarrollo de un modelo predictivo que determine el tiempo promedio de duración de una causa procesal en las materias no penales del año 2015 en el Consejo de la Judicatura.

1.6.12 Hipótesis

Si se verifica la existencia de la relación entre los propios registros de los procesos judiciales, entonces se determina un patrón de comportamiento en las demandas judiciales.

1.6.13 Hipótesis nula

Si se verifica la no existencia de una relación en los registros de los procesos judiciales, entonces no se determina un patrón de comportamiento en las demandas judiciales.

CAPÍTULO II
MARCO TEÓRICO, ANÁLISIS CONCEPTUAL Y ESTADO DEL ARTE

3.2.3 Marco teórico

3.3.1 Antecedentes Históricos de la minería de datos

Durante los inicios del año sesenta, se tiene indicio que, mediante conferencias y congresos, los estadísticos utilizaban los términos de data archeology, data mining o data fishing para referirse al análisis de los datos, el cual consistía, en aquellos tiempos la exploración de los datos para encontrar patrones que puedan llevar a descubrimiento de conocimiento de interés para el negocio. Pero no fue hasta los años ochenta cuando un grupo de investigadores comenzaron a unificar y relacionar los términos de KDD y minería de datos. Este grupo de investigadores estaba conformado por GioWiederhold, Rakesh Agrawat, Robert Blum, Piatetsky-Shapiro, Greogory entre otros.

Sin embargo, durante la misma época, grandes grupos de personas y empresas se mostró interesada por esta tecnología debido a sus aplicaciones, conformen han pasado los años esta tecnología ha seguido influenciada poco a poco hasta adentrarse en el corazón de los centros de investigación de mercado del consumidor y ha seguido expandiéndose a distintas áreas y en todo el globo terrestre. Si bien es cierto, grandes son sus aplicaciones, pero también grande son las confusiones que existen debido a que se la relacionan con simples reportes de distinta clase, lo cual es totalmente falso porque en esta área está inmersa la estadística, la cual sirve de base para los algoritmos que se implementan sobre ella para predecir eventos, obtener perfiles y patrones de consumo entre otros.

En la actualidad se tienen muchas herramientas que pueden ayudar a exprimir los datos para obtener conocimiento, pero para lograr ese conocimiento tan codiciado se debe seguir un proceso que permita llegar a él y esto se logra con el proceso KDD

(Knowledge Data Discovery); este utiliza algoritmos de minería de datos para obtener conocimiento a partir de datos prefabricados, esto significa que los datos pasan por un proceso de transformación.

3.3.2 Antecedentes conceptuales y referenciales de la Minería de Datos

Base de Datos: Según Byers, (1986). Es un conjunto de información organizada y presentada para servir a un propósito específico. El conjunto de datos es organizado en Tabla s, donde las Tabla s contienes filas y columnas; las filas se denominan registros y las columnas campos. Cada campo contiene determinado tipo de datos y tiene una longitud expresada en el número de caracteres máximo del campo.

Knowledge discovery in databases (KDD): "El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia Artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos)" (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996).

Business Intelligence (BI): Según Howard Dresner (1989) considera a la inteligencia de negocios como conjunto de conceptos y métodos para mejorar la toma de decisiones en los negocios, utilizando sistema de apoyo basados en hecho.

Minería de Datos (DM): Según Hand, (1998). Considera a la Minería de Dato como el proceso de análisis secundario de grandes bases de datos destinada en la búsqueda de relaciones insospechadas que son de interés o valor de los propietarios de la información.

CRISP-DM: Shearer menciona que: "Se trata de un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en minería de datos" (C, 2000).

3.2.4 Estado del Arte

3.3.3 ¿Qué es Minería de Datos?

De acuerdo a los conceptos de Fayyad (1996) y Molina (2001) definen que se denomina minería de datos (data mining) al análisis de archivos y bitácoras de transacciones, con el objetivo de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones. Desde mi punto de vista, considero a la minería de datos como un subproceso del KDD, el cual permite obtener modelos y/o patrones a través de los datos de las bases transaccionales o data warehouse, que a su vez ser utilizados en distintas ramas de la ciencia.

Aplicaciones de Minería de Datos.

En la actualidad, la minería de datos ha ido influenciado a todas las empresas del mundo, tanto es así, que cada año se realizan congreso en Estados Unidos donde los investigadores comparten sus trabajos y experiencias sobre las distintas ramas de la ciencia. También, se debe destacar el rol de las empresas, gobiernos y universidades en su apoyo por profundizar en esta rama la investigación.

Sin duda, todos de alguna u otra forma son recompensados por sus esfuerzos porque se sabe que las empresas aplican la minería de datos para explorar su información y obtener beneficios mientras que los gobiernos la utilizan para reducir la delincuencia analizando tendencias de robos y las universidades la aplican en productos tecnológicos para mejorar la calidad de las personas. Luis Carlos Molina (2002) en su investigación hace énfasis en las distintas aplicaciones de la minería de datos, pero en esta investigación se enfocará en dos de ellas que son en el entorno: gobierno y empresa.

En el estado La Policía Federal (FBI) de los Estados Unidos de América examinará las bases de datos de los negocios para detectar amenazas de extremistas

A mediados del año 2002, Jhon Aschcroft director del FBI realizó un comunicado en el cual manifestaba que el Departamento de Justicia de Estados Unidos de América comenzará a analizar la información comercial de todo el país con el objetivo de obtener patrones de compra de los clientes y de esta forma descubrir y prevenir posibles ataques de extremistas. No faltaron las personas que les asusto esta declaración por parte del director del FBI debido a que se está atentando contra la privacidad de información de las personas. Muchos expertos opinaron que, con el acceso a nivel nacional de todas las bases de los negocios, podrán saber los hábitos de comprar de un consumidor.

En las organizaciones Las empresas comenzarán a analizar las transacciones de las tarjetas de crédito para prevención y detección de fraude.

Expertos en la materia de fraude electrónico consideran que, en el año 2001, fue uno de los más tormentosos para las instituciones financieras de todo el mundo, debido a que miles de tarjetas de crédito fueron utilizadas de forma ilegal para sustraer aproximadamente dos mil millones de dólares. Y debido a todas estas pérdidas surge la necesidad de implementar un sistema antifraude llamado Falcon Fraud Manager, el cual analiza las transacciones de los movimientos bancarios del individuo con el objetivo de detectar conductas atípicas y de esta forma prevenir los fraudes que puedan ser realizados por piratas.

En la actualidad, el sistema Falcon Fraud Manager ha sido actualizado respecto a sus mejorar, incorporando nuevas funcionalidades al análisis de tarjetas comerciales, tal como se mencionó en el párrafo anterior, cuyo objetivo es mejorar el servicio de prevención de fraude. Este sistema ha permitido a los bancos ahorrar cientos de millones de dólares en pagos por fraude a sus clientes y, asimismo, se encarga de proteger las actividades realizadas con las tarjetas de crédito.

En conclusiones, se ha demostrado mediante ejemplos las distintas aplicaciones de la minería de datos en distintas áreas y en muchas de ellas compartiendo un factor común, el cual consiste en obtener conocimiento para beneficios propios de las

empresas. También hay que mencionar que los algoritmos utilizados por una empresa no pueden cumplir el mismo objetivo que otra debido a que las reglas de negocios son diferentes.

3.3.4 Tipos de Datos

Discretos.- De acuerdo a Triola Mario (2004) considera a los datos discretos resultan cuando el número de posibles valores es un número finito, o bien, un número que puede contarse. Por ejemplo: 0, 1,2, etcétera. (p.7). En otras palabras, los valores de una columna de atributos discreta no pueden implicar la ordenación, aun cuando los valores sean numéricos. Los códigos telefónicos de cada zona son un buen ejemplo de datos numéricos discretos.

Continuo. - "Datos continuos (numéricos) resultan de un infinito de posibles valores que pueden asociarse a puntos de alguna escala continua, cubriendo un rango de valores sin huecos ni interrupciones" (Triola Mario, 2004, pag.7).Es decir que los datos pueden contener un número infinito de valores fraccionarios.

3.3.5 El proceso KDD (Knowledge Data Discovery)

El proceso de KDD consta de cinco fases diferenciadas (José Hernández Orallo, 2004) , en este proyecto se centrará principalmente en una de ellas, la de minería de datos, pero también se mencionarán las demás, de cómo se seleccionan, limpian, transforman y almacenan estos datos. Y por supuesto una vez finalizada la etapa de la minería se tendrá que saber cómo interpretar y evaluar las conclusiones obtenidas de la investigación. A continuación, se describirán cada una de las fases para tener un breve conocimiento sobre cada una de ellas.

Fase de recopilación e Integración.- Esta fase se dedica a recopilar la información necesaria que sea de utilidad para el proceso de extracción de conocimiento y ellas pueden ser:

- Base de datos: pueden ser Data Warehouse o transaccionales

- Archivos: los archivos pueden ser .txt, csv, json, xml entre otros

Fase de selección, limpieza y transformación.- En esta fase se limpian los datos erróneos, incompletos o inconsistentes (limpieza) e irrelevantes (criba).Además, de la limpieza Francisco Barrios y Sebastián Ríos describen que pueden generarse nuevas variables a partir del estudio de la naturaleza de las variables originales. (Barrientos & Ríos, 2013)

Algunos métodos imprescindibles para la limpieza de datos:

- Histogramas (visualizar distribución de los datos).
- Diagrama de cajas (visualizar los datos atípicos de un conjunto de datos)
- Selección de datos (conjunto de datos a ser seleccionado para el análisis)
- Agrupación o separación de datos (se realiza esto en caso de ser necesario)

Actividades a realizar antes datos atípicos:

- Analizar los datos: se realizan consultas para verificar que cantidad de registros representan los datos atípicos.
- Discriminación: Se los datos atípicos representan una gran afectación a tu modelo y la representación de estos es menor al 5 % es mejor ignorarlos. Pero esto depende del proyecto que realices. Por ejemplo, si se tratase de fraudes de movimiento bancarios esto no debería ser obviado.
- Realizar diagrama de caja: los diagramas de cajas son excelentes para visualizar todo un conjunto de datos y ver si existen datos atípicos.
- Reemplazar de datos: puedes aplicar el reemplazo datos si tienes un campo género, pero sucede que los registros tienes masculino, hombre y debes llevar todo a un solo nombre. En estos casos se suele actualizar la columna con un solo nombre y esa decisión la tiene el investigador.

Actividades a realizar ante datos faltantes:

- Omitir: puedes omitir los datos faltantes si no representan utilidad para la investigación.
- Segmentar: Si los datos faltantes son un porcentaje pequeño de tus datos, lo mejor es trabajar en base a la segmentación y de allí generalizar los resultados para todos los datos.
- Reconstruir valores: si los datos faltantes representan un gran porcentaje y son claves para la predicción se debe aplicar algoritmo que generen esos valores en base a promedios. Es decir, en base a tus datos existentes, se deben regenerar los nuevos valores.

Fase de Minería de Datos.- De acuerdo a Francisco Barrientos y Sebastián Ríos mencionan que: "En esta parte del proceso KDD, consiste en la aplicación de análisis de datos para descubrir un algoritmo ad-hoc que produzca una particular enumeración de patrones a partir de los datos" (Barrientos & Ríos, 2013). Por lo tanto, esta fase se basa en descubrir patrones que pueden servir de utilidad y para ello se debe tener presente lo siguiente:

- Definir los objetivos de la minería de datos y verificar que sean medibles y alcanzables.
- La técnica de minería a ser aplicada va en relación al conocimiento que se quiera obtener y esto debe térnelo claro el explorador.
- Desde el inicio se debe tener claros los objetivos a descubrir porque sin ellos toca torturar a los datos hasta encontrar alguna relación.
- En algunos casos cuando no se tiene idea de lo que se desea encontrar es una buena práctica realizar una matriz de correlación eso es como una radiografía de los datos.

Fase de evaluación y validación.- Una vez obtenido la o las hipótesis de los objetivos a descubrir de la fase anterior, se procede a seleccionar y validar los mismos basados en criterios de evaluación relacionados a los fines que persigue la institución.

A continuación, se describen algunas actividades que se deben tener presentes para esta fase:

- Comparar con uno o más modelos los resultados obtenidos para seleccionar el modelo que más se acerca a los objetivos perseguidos.
- Verificar que la precisión del modelo sea la idónea.
- Realizar las pruebas de tus modelos con un 40% de tus datos, pero si la cantidad de registros es pequeña se puede optar por utilizar todos los registros para las pruebas.

Fase de interpretación y difusión.-En esta última fase se verifican e identifican patrones obtenidos que son de interés para luego realizar una trasposición de los resultados técnicos a niveles comerciales beneficiosos para el negocio. (Barrientos & Ríos, 2013) Algunas consideraciones que se deben tener en esta fase se describen a continuación:

- Si se tiene la aprobación de implementar el modelo se tiene que realizar un plan de implementación que mencione las actividades a realizar. Puede que la implementación requiera advertir a los usuarios por celular o aplicativo web o tener acceso directo a la base de datos para informar en tiempo real. Todos estos aspectos se deben tener en cuenta en plan que se genere.
- Algunas veces, los modelos necesitaran ser interpretados debido a su complejidad de los resultados. Por ejemplo, un modelo de árbol que me indica todas las condiciones que pueden pasar dependiendo del sector donde vivas y el lugar de donde trabajes y al igual que el primer paso se necesita de un plan difundir e interpretar el conocimiento.

3.3.6 Técnicas de Minería de Datos

Existen dos clasificaciones o técnicas en la minería de datos tales como: no supervisados y supervisados. Los algoritmos no supervisados son aquellos no

necesitan de la intervención humana para interpretar resultados porque al pasar por un proceso riguroso de evaluación su respuesta es siempre satisfactoria. Mientras que los algoritmos supervisados son muy dinámicos, es decir tienen entradas de datos cambiantes por lo cual el resultado va en función de dichas entradas, lo que significa que estos deben adaptarse a las nuevas variables para obtener nuevos resultados.

En la Tabla 1, siguiente se muestran algunas de las técnicas de minería de ambas categorías.

Tabla 1
Clasificación de las técnicas de minería de datos.

SUPERVISADOS	NO SUPERVISADOS
Inducción Neuronal	Segmentación
Árboles de decisión	Agrupamiento(“clustering”)
Regresión	Reglas de asociación

Modelo agrupamiento (K-Means).- Esta técnica de minería de datos no supervisada divide un conjunto de datos en subgrupos llamados clases que a su vez se diferencian por su máxima distancia de separación entre ellas. Se considera una clase cuando todos sus elementos tienen la mínima separación de distancia posible entre ellos.

Reglas de asociación.- Según Haydeé Gommez y María de los Angeles Cerón resaltan que: “El método de asociación detecta eventos que ocurren de manera simultánea” (Gommez Díaz & Cerón Reyes, 2010). El objetivo de esta técnica es encontrar patrones de asaciones que se relacionen de alguna forma o que den pistas de nuevas relaciones (Morales & Gonzáles, 2012). Las reglas de asociación tienen distintas aplicaciones como, por ejemplo:

- Encontrar las relaciones de productos que más se venden con el objetivo de formar canastas de alimentos o promociones.

- Obtener perfiles de consumidores

(Berzal, 2013) El proceso de evaluación de las reglas de asociación en minería de datos se realiza de acuerdo a:

- La cobertura (soporte): Es el número de instancias para las cuales ella predice correctamente (soporte).
- La precisión (confianza): Es el número de instancias que predice correctamente, expresado como una proporción de todas las instancias a las que se aplica.

Modelo de red neuronal.- Es una técnica de minería de datos predictiva que imita las redes de inteligencia del cerebro aprendiendo de sí mismas bajo entrenamiento y estas se pueden clasificar en supervisadas y no supervisadas. Se podría decir que la red neuronal son cajas negras, porque no siempre siguen un proceso lógico o comprensible para el ser humano. Sin embargo, su interés radica en que son herramientas útiles para realizar predicciones, por lo que son usadas en numerosas aplicaciones.

Modelo de regresión lineal.- Es una técnica predictiva netamente estadística utilizada en la minería de datos que estudia el cambio de la variable independiente en relación de la variable dependiente (Moral Peláez, 2012).

Modelo de árbol de decisión.- Es una técnica que ayuda a la toma de decisiones y su función consiste en segmentar el problema en subclasificaciones de decisiones sucesivas en donde cada rama del árbol tiene una probabilidad. Estos árboles se dividen en: árboles de regresión y clasificación.

3.2.5 La metodología CRISP-DM

IBM resalta que: "CRISP-DM son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar los trabajos de minería de datos" (IBM, 2012). Cuando se trata de implementar un proyecto de minería de

datos, usualmente se recurre de forma subjetiva a nuestra experiencia o se asocia a técnicas y métodos utilizado en otras ramas que pueda ser de utilidad pero cuando el proyecto es de gran magnitud y se vuelve complejo nuestras herramientas tienden a fallar y es así que IBM sugiere que para evitar contratiempos y problemas a futuro en proyecto de minería de datos se deba utilizar CRISP-DM porque considera una metodología probada y utilizada por empresas de gran envergadura.

Para comprobar las palabras de IBM se realizó una búsqueda en kdnuggets.com (sitio web orientado a la ciencia de los datos) sobre la metodología más utilizada y para nuestra sorpresa esta es CRISP-DM. El resultado lo puedes apreciar en la Figura 1.

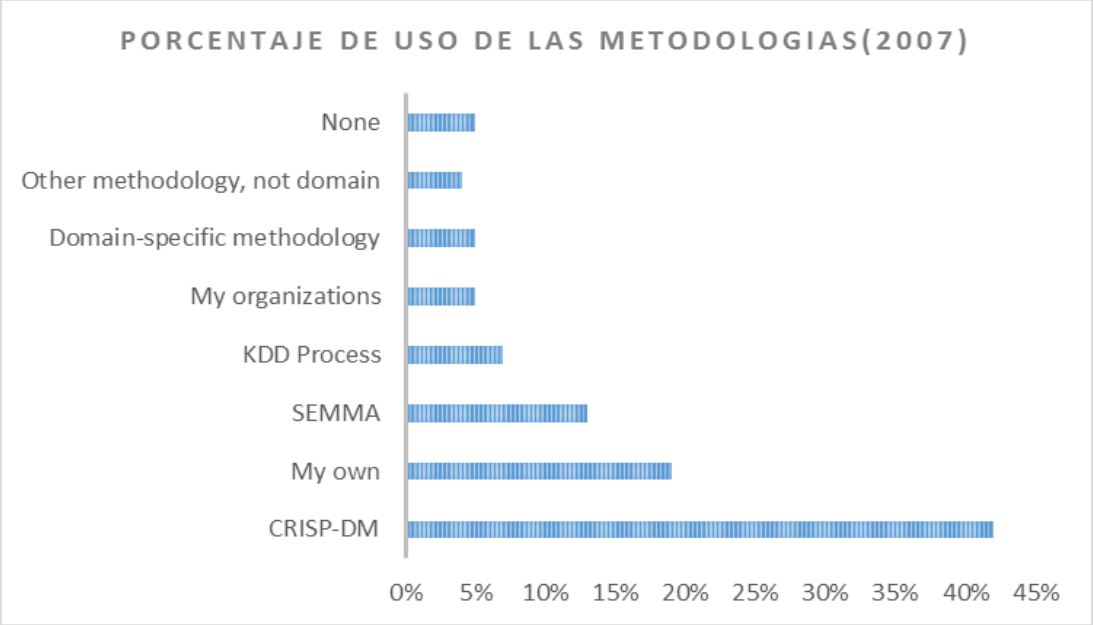

Figura 1. Metodología más utilizada en el año 2007.
Fuente: (KDnuggets, 2016).

La metodología CRISP-DM, tal como se muestra en la Figura 2, se compone de 6 fases bidireccionales que indican la secuencia que se deben realizar para alcanzar los objetivos. Estas fases son bidireccionales porque permiten avanzar o regresar hacia una fase en caso de que este inconcreta o se haya definido mal los objetivos. Se puede decir que es una metodología muy flexible pensada y orientada a los cambios que puedan suceder.

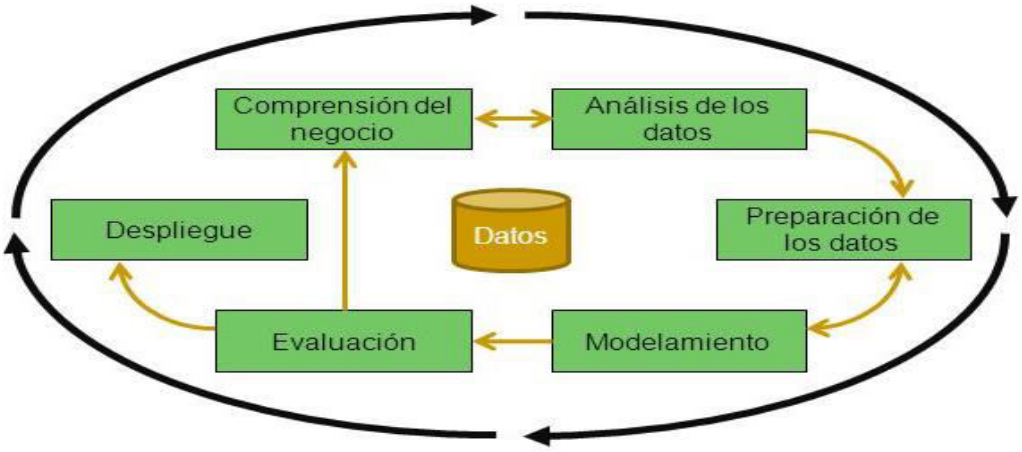

Figura 2. Modelo de proceso CRISP-DM.
Fuente: (Rojas, 2010).

A continuación, se detalla la descripción de cada una de las fases de esta metodología.

1. **Fase de comprensión del negocio o problema**

Marcelo Barrios resalta que: "en esta fase se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el plan de trabajo" (Barrios, 2010). Esta fase es de extrema importancia debido a que el explorador debe mentalizar y comprender el negocio para poder definir junto con la empresa los objetivos de la minería. Como se mencionó con anterioridad, los objetivos deben ser medibles y alcanzables a corto plazo porque eso busca una empresa, resolver sus problemas en el menor tiempo posible.

Si por alguna anomalía los objetivos no fueron definidos correctamente, entonces no se sabe lo que se busca y así aplique miles de algoritmos no sabrá que interpretar porque simplemente no se tienen claro lo que se va a hacer. En la Figura 3 se muestra el flujo del proceso.

Figura 3. Fase de comprensión del negocio.

Fuente: (Pete Chapman, 2000).

Dentro de esta fase existen ciertas actividades que se deben realizar y se las menciona a continuación:

- **Determinar los objetivos del negocio.**

Como una de las primeras actividades dentro de la fase de comprensión del negocio se debe tener claro que problema desea resolver el dueño de la empresa. En estos casos es necesario platicar con los involucrados para tener una idea clara de lo que se debe hacer luego de eso se debe pasar a definir los objetivo. Este proceso es parecido a la obtención de requisitos, donde las dos partes se reúnen para retroalimentarse mutuamente. (CEI, 2010). Como parte relevante de este proceso, esta tarea tiene una enorme importancia debido a que describe el problema que se desea solventar. Si esta tarea falla en definir los objetivos que persigue la entidad, todo el proceso se verá involucrado.

- **Evaluación de la situación**

Una vez determinados los objetivos se pasa a evaluarlos para saber si se tiene la información necesaria para poder resolverlos. Puede que la empresa quiera alcanzar

dichos objetivos, pero sucede que no tiene información por lo tanto dicho objetivo será puesto a un lado.

- **Determinar los objetivos de la minería de datos**

Una vez determinados y evaluados los objetivos del negocio se procede a definir los objetivos de la minería los cuales deben ser cortos en términos de tiempo y medibles que se pueda obtener resultados. Un ejemplo para tener una idea clara de cómo definir un objetivo sería "Definir los productos de la canasta básica para las compras del mes de diciembre". Obviamente saltaran ciertos preguntas, a que sectores del país, o provincias. Todas estas dudas deben ser aclaradas durante la definición del objetivo.

- **Realizar el plan del proyecto**

En esta actividad se debe realizar un documento con las instrucciones a seguir en las siguientes fases. Aunque analizando la metodología, esta indicación esta demás debido a que CRISP-DM ya indica los pasos a seguir.

2. Comprensión de los datos

En esta fase se debe recolectar los datos que sean de utilidad para nuestra investigación, en otras palabras, irían en relación a los objetivos definidos, luego se procede a describirlos para tener una idea clara de los datos para pasar posteriormente a explorarlos utilizando herramientas o técnicas estadistas para ver su distribución, comportamiento entre otros aspectos. (Folgueiras Bertomeu, 2010).

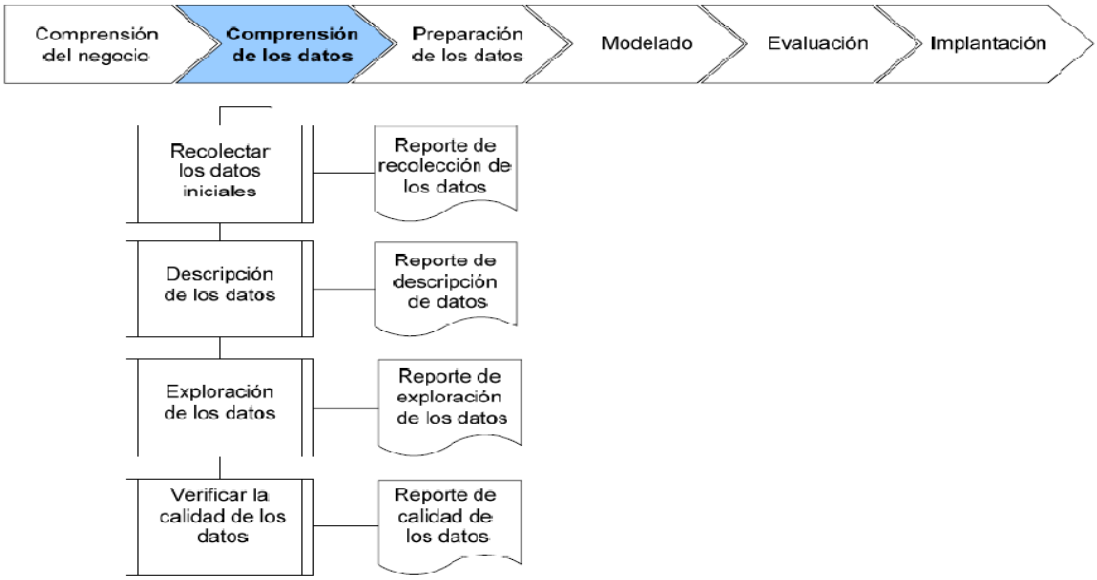

Figura 4. Fase de comprensión de datos.
Fuente: ***(Pete Chapman, 2000)*.**

En la Figura 4, se observa las actividades que se deben realizar dentro de esta fase, a continuación, se las describe:

- **Recolectar datos iniciales**

Antes de realizar la recolección de los datos, se recomienda revisar primero los objetivos de la minería de datos para en base a esa relación solicitar los datos. Si usted evita este pasó, puede que pida toda una base de datos que sea imposible analizar por los tiempos y afectaría a la ejecución del proyecto. (CEI, 2010)

- **Descripción de los datos**

Se recomienda en esta actividad realizar una descripción general de los datos que describa los atributos de los campos, formatos, Tabla s involucradas entre otros. (Barrios, 2010)

- **Exploración de los datos**

En esta sección de exploración de datos se tiene que sumergir dentro de ellos para encontrar anomalías, herramientas poderosas para realizar radiografías de los datos

son los histogramas que permiten ver la distribución de los datos para tener una idea general y los diagramas de cajas que permiten detectar datos atípicos. Y existen muchas otras técnicas que puede aplicar usted dependiente de lo que busca.

- **Verificar la calidad de los datos**

Lo que se debe realizar en esa actividad es un examen completo de los datos dentro de cada atributo y tratar de encontrar defectos y cuantificarlos para tener porcentajes sobre la cantidad que representan de mi conjunto. (Folgueiras Bertomeu, 2010).

3. Preparación de los datos

Del conjunto de datos proporcionados por el departamento de tecnología, se debe seleccionar que parte de ellos se utilizaran para el análisis y luego estos pasaran a ser limpiados, por ejemplo si se tiene en un campo fecha valores como "00-00-00" estos deberán ser limpiado o ignorados.

El siguiente paso a seguir es armar la estructura donde se almacenará el resultado y posteriormente unificarlos. Algunas veces se tiene a colocar los datos en Tabla s desnormalizadas para evitar tener tantas Tabla s. (Rigeiro, 2012). La siguiente imagen explica mejor las actividades internas de esta fase.

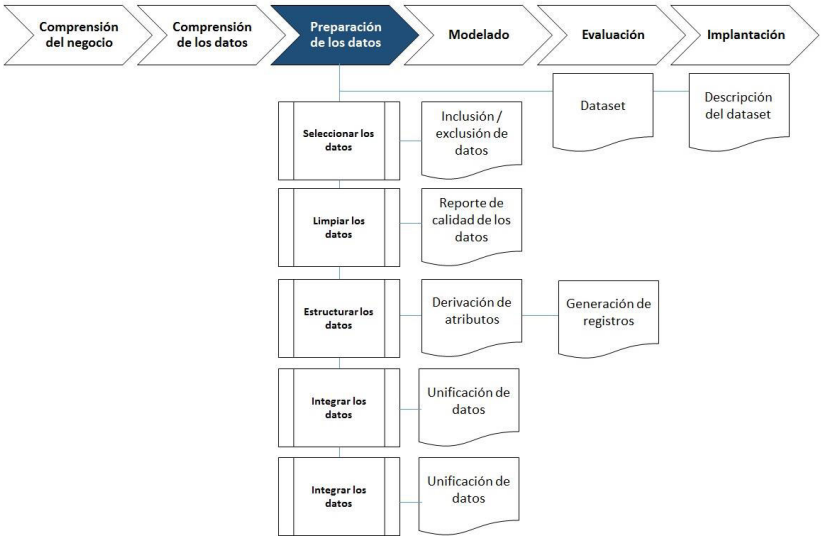

Figura 5. Fase de preparación de los datos.
Fuente: ***(Pete Chapman, 2000)*****.**

- **Seleccionar los datos**

Consiste en seleccionar un conjunto de datos del dataset que tengan todos los datos a ser analizados completos.

- **Limpiar los datos**

La limpieza de los datos es muy especial y tiene mucha importancia debido a que si los datos presentan problemas de integridad sencillamente los algoritmos no funcionaran y no se podrá llegar a ningún resultado.

- **Construir los datos**

José Luis Llavona Arregui destaca que: “En esta fase se lleva a cabo la construcción de nuevos datos, derivados de los disponibles, que son importantes para el análisis."(Llavona Arregui, 2010)

Esto significa que puedes crear otros campos a partir de los ya existentes, todo va en relación a la necesidad que se presente.

- **Integrar los datos**

En esta sección se trata de unificar todos los datos de las distintas Tabla s para obtener una Tabla resumen (desnormalizada) en la cual se hará más sencillo realizar el análisis.

- **Formateo de los datos**

Consiste en eliminar caracteres basura dentro de los registros ya que en algunas bases de datos los campos suelen tener tabulaciones al final de la línea y esto representa un gran problema si lo que se quiere analizar en una variable categórica.

4. Modelado

En esta sección ya debe tener claros los objetivos, haberse involucrado en el problema de la empresa, comprender el negocio y disponer de la información a analizar. Si cumple con los requisitos, la selección del modelo no es un problema.

Pero como todo modelo no es perfecto, se necesita de una evaluación y justamente durante esta fase de realiza ese proceso que consiste en comprar dos o más modelos y ver cuál de ellos entregan mejores resultados o que más se ajuste a las necesidades de la investigación. En la siguiente Figura 6 de CRISP-DM indica las actividades a realizar.

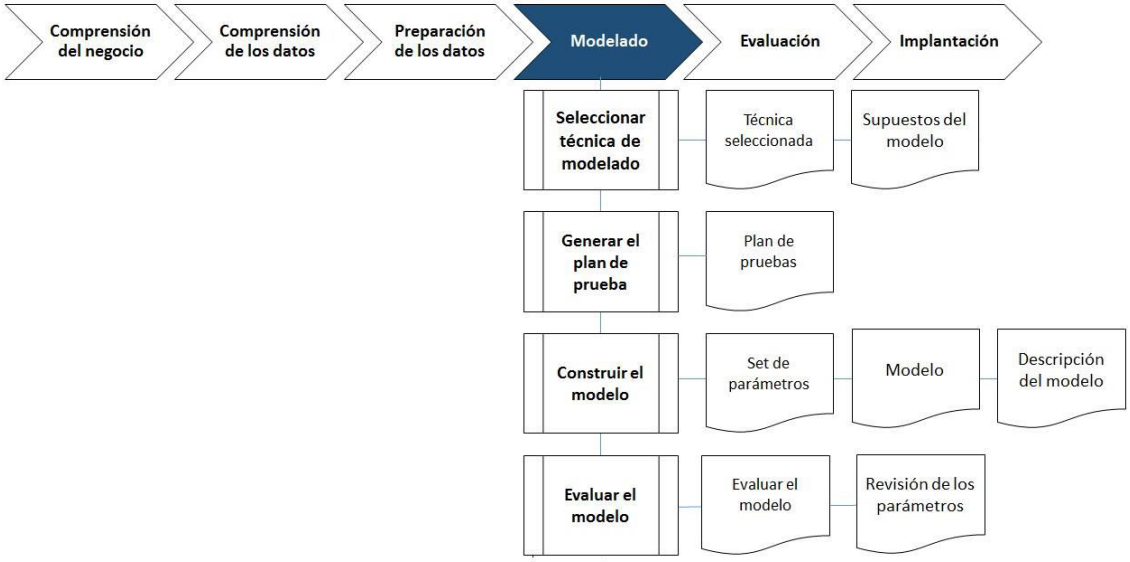

Figura 6. Fase de modelado.
Fuente: ***(Pete Chapman, 2000)*****.**

- **Escoger la técnica del modelado**

La selección de la técnica no presenta problema lo que sí es el resultado que se obtenga. Por eso en esta actividad hay que probar con distintos modelos. Por ejemplo, se puede probar con una regresión lineal, pero sucede que los resultados no son exactos pero si aplica un modelo SVM radial sucede que los resultados son muy aproximados entonces se procede a seleccionar esa técnica. (Molero Castillo, 2008)

- **Generar el plan de prueba**

Como se mencionó al inicio de esta fase, se debe probar el modelo para medir su fiabilidad y tasas de error que se obtiene durante los resultados. En la actualidad, muchas herramientas traen modelos que ya cuentan con la tasa de error incorporado, cuando se entrenan sobre datos, pues estos deberán ser tomados en apuntes para más adelantes hacer la comparación del mejor modelo a seleccionar. En resumen, esta

sección lo que hace es medir la tasa de error de cada uno de los modelos con el fin de seleccionar el modelo más idóneo.

- **Construir el modelo**

En esta actividad se ejecutan los modelos con los datos seleccionados y el investigador debe seleccionar cuáles de ellos pasaran a proceso de evaluación. (Molero Castillo, 2008)

- **Evaluar el modelo**

Después de haber ejecutado los modelos en la fase de construcción, en esta parte se analizan los que han sido seleccionados en base a su fiabilidad y la menor tasa de error generado. El investigador tendrá que seleccionar en función de los objetivos de la minería el modelo que más se ajuste y cumpla con las expectativas.

5. Evaluación

De acuerdo a la metodología CRISP-DM se tienen que evaluar los resultados, pero de forma orientado al negocio y se debe realizar un proceso de revisión para determinar si los modelos obtenidos están orientados a solventar los objetivos definidos por la empresa. Si el modelo seleccionado (el mejor) cumple con los objetivos y tiene una fiabilidad determinada, producto de la comparación de los modelos, se procede a su uso en la entidad. Observar la Figura 7 para conocer las actividades que se realizan dentro de esta fase.

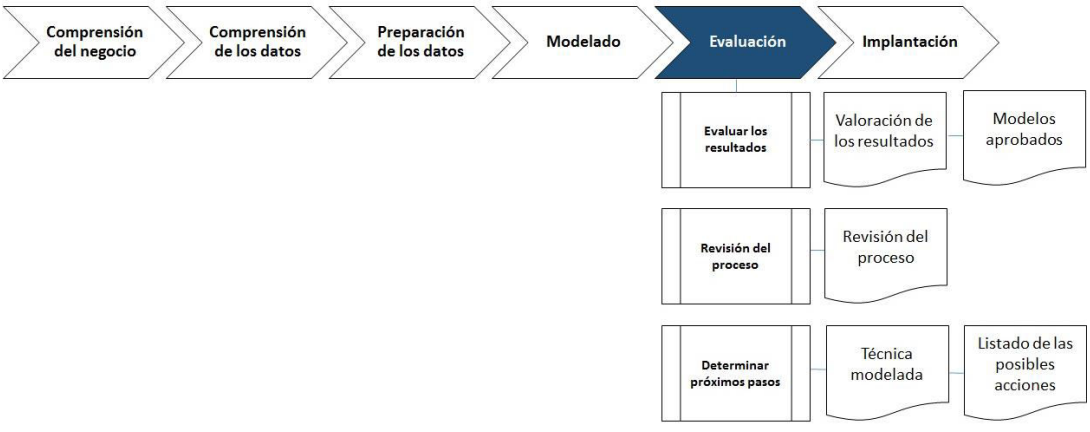

Figura 7. Evaluación del modelo.
Fuente: ***(Pete Chapman, 2000)*****.**

- **Evaluar los resultados**

Tal como se describió al inicio de esta fase, los objetivos deben ser evaluados en relación al negocio que fueron definidos en la fase inicial de la metodología. Aquí se toman los modelos resultantes y el equipo tendrá que seleccionar la opción que cumpla con su necesidad empresarial.

- **Revisar el proceso**

Durante esta actividad se efectúa una revisión de todo el proceso realizado si hubiese procesos que necesiten optimización se procede a realizarlos.

- **Determinar los próximos pasos**

Esta tarea es clave para avanzar o retroceder debido a que si los resultados son incongruentes con los objetivos se tendrá que redefinirlos, pero de caso contrario se procederá a su implantación.

6. Despliegue o implantación

CRISP-DM sugiere que en esta fase se debe implementar el modelo seleccionado, probado y evaluado para su puesta en ejecución en la empresa y de esta manera forme parte de la lógica de negocios de la entidad. Pero esto raramente ocurre porque los resultados tienen que pasar por fases de pruebas, pero no de entrenamiento si no de los datos reales. Además, los datos deben ser socializados con los clientes.Tareas de esta fase se exponen en la Figura 8.

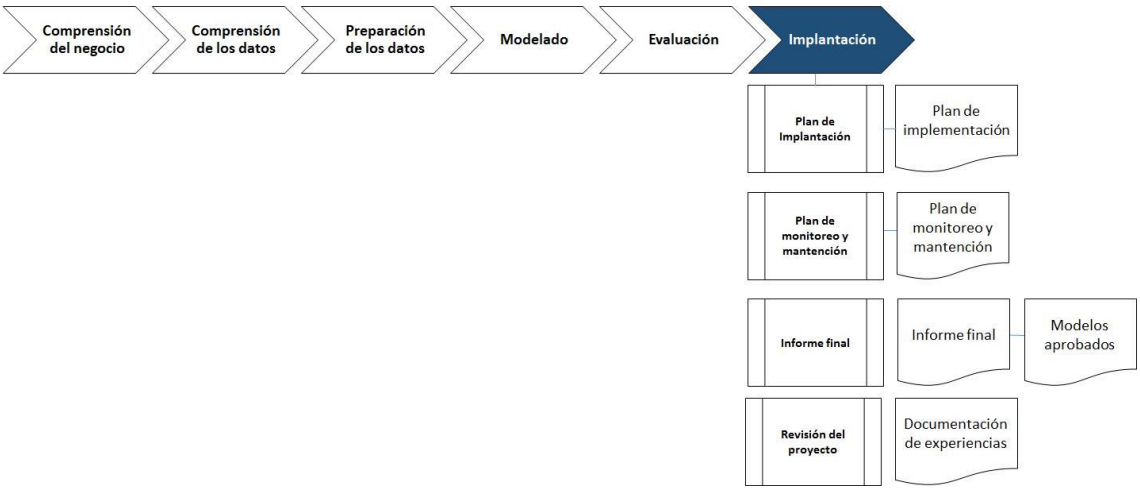

Figura 8. Fase de implantación.
Fuente: ***(Pete Chapman, 2000)*****.**

- **Planear la implantación**

En este punto se procede al desarrollo de un plan de implementación donde se definas que instrucciones se debe seguir para la utilización del modelo en la empresa. (Moine, Haedo, & & Gordillo, 2012)

- **Planear la monitorización y mantenimiento**

En esta fase se debe definir un plan de contingencia para la monitorización y mantenimiento del modelo para saber cómo actuar ante situaciones atípicas.

- **Producir el informe final**

En esta parte, el equipo desarrolla un informe donde evidencia los resultados obtenidos durante la realización del proyecto. En conclusión, podría ser un informe o una diapositiva que trate los puntos importantes del proyecto explicando los resultados logrados en el mismo.

- **Revisar el proyecto**

Como etapa final de la metodología tiene la opción de permitirle regresar al investigador a cualquier fase para corrección en caso de tener incongruencias con algún objetivo o mal interpretación de datos en cualquiera de las fases

CAPÍTULO III
METODOLOGÍA

3.3.7 3.1 Uso de la Metodología CRISP-DM

Entrando ya en la parte práctica del proyecto donde se ejecutarán cada una de las fases de la metodología, se concentrará en la resolución del problema que se ha planteado para explotar los datos de la función judicial. A continuación, se inicia con la comprensión del negocio de la función judicial.

3.3.8 3.2 Comprensión del negocio

La función judicial como ente principal del estado ecuatoriano, es la encargada de administrar la justicia, es decir, ofrecer las herramientas a los usuarios para que puedan tramitar sus denuncias, realizar sus seguimientos, contactar con las autoridades para ejecutar órdenes de detención si fuesen necesarias, solicitar pensiones de alimentos, realización de allanamientos, providencias preventivas entre otras. Todo es posible gracias a la meta que tiene la función judicial, de ofrecer el mejor servicio de justicia a sus ciudadanos. De acuerdo a la metodología CRISP-DM tocaría iniciar con el proceso de definir los objetivos del negocio en la función judicial.

3.3.9 3.2.1 Objetivo del negocio

El objetivo es desarrollar un modelo predictivo mediante el análisis de tendencias y patrones de duración de las demandas judiciales para determinar el tiempo promedio de duración de un juicio.

3.3.10 3.2.2 Evaluación del negocio

La función judicial es consciente de la problemática que existe hoy en día, los ciudadanos desconocen el tiempo promedio de duración de un proceso judicial, lo que se refleja mayoritariamente a desistir de los procesos legales produciendo que no

sea práctico llevar un juicio ante la ley. Además, las personas asocian el tiempo con dinero, lo que significa a mayor tiempo se prolongue el juicio mayor dinero será el que se tenga que gastar. Adicionalmente se tiene que agregar otros factores que intervienen en el proceso tal como los abogados, los cuales aprovechan el desconocimiento de los ciudadanos respecto a la duración de una causa judicial para prolongar el tiempo y lucrarse de ello.

Teniendo en cuenta, la problemática mencionada en el párrafo anterior, otro punto que se tuvo en cuenta para el análisis, son los datos que son de utilidad para llevar a cabo el proceso de minería de datos. La institución cuenta con un repositorio que tiene información de nueve años aproximadamente, pero los datos que se analizarán durante esta investigación son del año 2015, los cuales son suficientes para demostrar nuestra hipótesis.

3.2.6 Objetivos de la minería de datos

- Verificar la existencia de la relación entre los propios registros de los procesos judiciales del año 2015 en materia de familia niñez y adolescencia.
- Formular un modelo predictivo de la duración promedio de un juicio a través del análisis de patrones de los procesos judiciales del año 2015 en materia de familia niñez y adolescencia.
- Determinar las tendencias de los procesos judiciales utilizando el historial de registros de los procesos judiciales del año 2015 en materia de familia niñez y adolescencia.

El desarrollo del modelo predictivo de las demandas judiciales a través de la minería de datos puede ser de mucha utilidad a la hora de aplicar nuevas técnicas para mejorar el servicio de disponibilidad de la justicia a los ciudadanos. Todo esto permitirá mejorar la calidad de los servicios ofrecidos por la institución.

3.2.7 Realizar el Plan del Proyecto

En esta actividad se debe definir un pequeño cronograma con las actividades a realizar y los tiempos en cada una de ellas.

- Fase de comprensión del negocio: Reuniones informales con la parte de gestión procesal para entender el problema que tienen y, así mismo, se solicitará una explicación de las estructuras de almacenamiento de datos para poder tener una idea de que Tabla s se emplearán en la selección de datos. El tiempo empleado: 2 semanas.

- Fase de comprensión de los datos: Se analizará la información proporcionada, se verificará las relaciones entre las Tabla s y se obtendrá un conjunto de datos representativos. El tiempo empleado para esta tarea: 3 semanas.

- Fase de preparación de los datos: Se realizará limpieza de los registros, se realizarán conversiones de las fechas y se seleccionara el conjunto de datos a trabajar para el análisis. El tiempo para estas tareas es de: 6 semanas.

- Fase de modelado: Se seleccionará la técnica a ser empleada, se escribirá el código para la ejecución del modelo. Tiempo para esta tarea es de: 12 semanas.

- Fase de evaluación: Se examinará los resultados obtenidos de la fase de modelado y si se encontrará incongruencias se repetirá el proceso anterior. Tiempo para la tarea: 4 semanas.

- Fase de implementación: Se definirán pautas para la puesta en marcha del modelo en la institución. Tiempo estimado: 1 semana

- Presentación de los resultados: se realizará un informe resumido de los resultados de la investigación.

3.2.8 Evaluación de la técnica y selección de la herramienta

Para este caso como se está familiarizado con los programas open source se utilizará: R, R Studio y PostgreSQL ya que se adaptan bien a la metodología que se está empleando. Además, gracias a R Studio que es la interfaz para manejo de R, se facilita la lectura de los datos de la base de datos de PostgreSQL y así poder realizar el análisis de la información.

En cuanto a las técnicas que se van a emplear para la extracción de conocimiento, R ofrece paquetes con los siguientes paquetes:

- En las predictivas se encuentran las siguientes:
  - Regresión
  - Clasificación
- Respecto a las descriptivas se tiene:
  - Reglas de asociación
  - Agrupamiento

R utiliza una cantidad enorme de algoritmos para la minería de datos entre ellos se encuentran las más destacadas: algoritmos a priori, arboles de decisión, máquinas de vector de soporte y regresión lineal. Pero en caso se quiere predecir tendencias y se ajustaría más las regresiones lineales, pero también los modelos SVM pueden realizar regresiones. Por lo tanto, se seleccionará ambos modelos para el propósito.

3.3.11 3.3 Comprensión de datos

Según CRISP-DM, para poder comprender lo datos se debe realizar una recolección de la información orientada a los objetivos de minería, posteriormente

hay que describirlos y explorarlos para saber a qué tipo de datos se enfrentan. Y finalmente ver su calidad para dejarlos listos para la fase de selección.

3.3.12 Recolección de los datos

Se debe mencionar que los datos empleados para esta investigación son de los procesos judiciales que incluyen información como: juicio, materia, tipo de acción, fecha de ingreso, fecha de resolución y tiempo de duración de la causa. Hay que aclarar que los atributos reales son: el tiempo de la resolución de la causa; el resto de campos como el juicio, la materia y el tipo de acción son ficticios debido a la sensibilidad de la información.

La información solo estará disponible para el análisis de los datos. Con esta información se podrá obtener la tendencia en base a un modelo de regresión lineal y determinar en base al análisis el tiempo promedio de una causa procesal. El paso de la información hacia nuestra base de datos postgreSQL se la realizó a través de un script de inserción leyendo un archivo Excel, los scripts de carga pueden ser visualizados en el Anexo 2. A continuación en la Tabla 2, se presentan los nombres de las entidades (nombre ficticio) relacionadas con los datos obtenidos:

Tabla 2
Descripción de la Tabla desnormalizada.

CAMPO	DESCRIPCIÓN
Juicio	El campo juicio tiene un identificador único. Este campo mantiene relación con las Tabla s materia y tipo de acción. (ficticio)
Materia	El campo materia tiene se diferencia por tener identificador único. Además, toda materia está relacionada con un tipo de acción. (ficticio)
Tipo de Acción	Cada tipo de acción está identificado por un número. Además, todo tipo de acción está relacionado con un juicio. (ficticio)
Tiempo	Tiempo de resolución de un proceso judicial

Técnicamente hablando, los siguientes campos que serán útiles para el análisis son:

- IdMateria
- IdTipoAccion
- NumDia
- FechaIngreso

De los atributos mencionados, algunos de ellos sufrieron transformaciones tal como: “fechaIngreso” de cadena a formato fecha, lo mismo se aplicó para los campos “IdMateria” e “IdTipoAccion” de cadena a entero. El fin de esto es facilitar el trabajo a los modelos.

3.3.13 Describir los datos

Los datos se encuentran almacenados en una Tabla desnormalizada, ya que por motivos de confidencialidad no se pueden exponer el nombre de las Tabla s verdaderos. Además, la Tabla desnormalizada, llamada “proceso_judicial”. En la Figura 9 se puede visualizar el esquema de la Tabla desnormalizada. Para generar esta figura se ha utilizado la herramienta gratuita online “dbdesigner.net”.

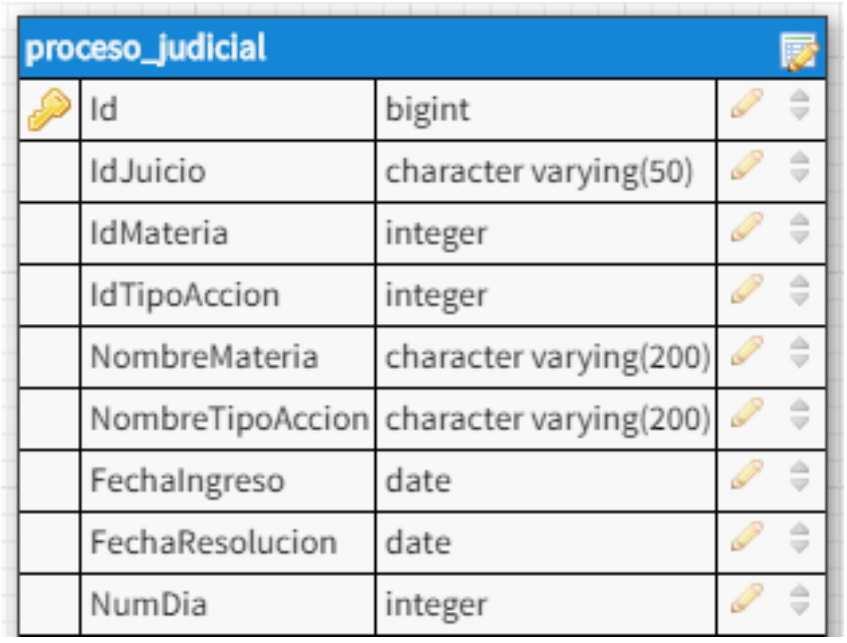

Figura 9. Esquema de la tabla de los procesos judiciales.

En esta Figura 9 se puede observar claramente los campos que conforman la tabla **proceso_judicial**. A continuación, se describirá cada uno de los campos que conforman la tabla.

Tabla proceso_judicial

Esta tabla es la tabla central, o también llamada "tabla de desnormalizada", es donde se registran los procesos judiciales terminados. Al ser información confidencial, el nombre de la tabla y el de los campos difieren totalmente de la realidad. El número de registros en esta Tabla es de 808.584 y está conformada por los siguientes atributos:

- **Id:** Es un campo auto numérico que se utiliza como identificado.
- **IdProcesoJudicial:** Tipo cadena. Este campo identifica al juicio y es único para cada proceso judicial. La generación del código del campo está conformada por 16 caracteres que son únicos.
- **IdMateria:** Tipo numérico. Este campo identifica a la materia del proceso judicial y es único para cada juicio. Las materias van desde no penales hasta penales.
- **IdTipoAccion:** Tipo numérico. Este campo identifica el tipo de acción de una materia en un proceso judicial. El tipo de acción es único para el juicio.
- **NombreMateria:** Tipo cadena. Este campo es el nombre de la materia dentro de un proceso judicial.
- **NombreTipoAccion:** Tipo cadena. Este campo es el nombre del tipo de acción de una materia en un proceso judicial.
- **FechaIngreso:** Tipo fecha. Este campo es la fecha de ingreso del juicio y es único por juicio. El formato en el cual se guardan los datos es: "yyyy-mm-dd".
- **FechaResolucionProceso:** Tipo fecha. Este campo es la fecha de culminación temporal del juicio. Esto significa que un juicio puede volver a ser reabierto por

apelación o por otros factores, pero se considera resuelto. El formato en el cual se guardan los datos es: “yyyy-mm-dd”.

- **NumDia:** Tipo numérico. Este campo almacena el total de días que se ha demorado en resolver un proceso judicial. Ejemplo: 100.

3.3.14 Explorar los datos

En esta tarea se pasará a explorar los datos mediante el uso de técnicas estadísticas implementadas en R. Se comenzará realizando histogramas de los juicios resueltos por meses de todo el año 2015, luego se realizarán tendencias mensuales para ver su proyección y finalmente se realizará graficas anuales. (Ver Anexo 2. Literal A. Scripts de las gráficas)

Histograma

En la Figura 10, se observa los histogramas mensuales del año 2015 de la materia familia niñez y adolescencia. En la gráfica de enero se observa que en los primeros cinco días se ha resuelto más casos judiciales que en el resto de días.

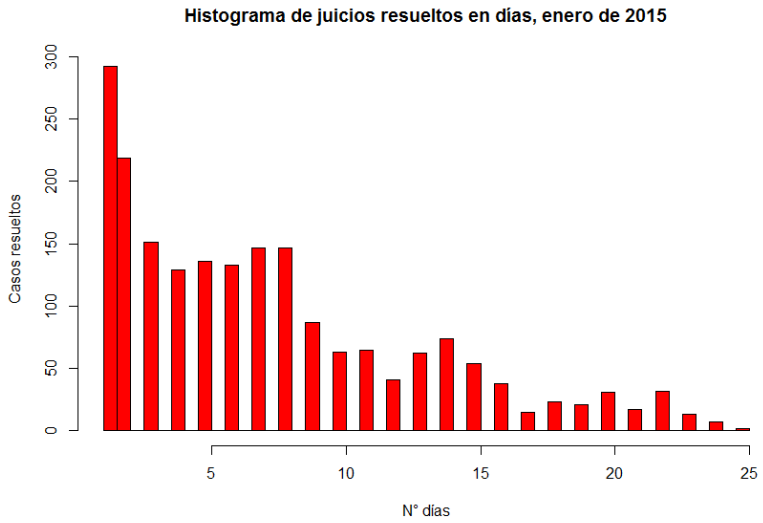

Figura 10. Histograma de juicios resueltos en días, enero de 2015.

En la Figura 11, se presenta el histograma del mes de febrero de 2015. Se observa que al igual que enero, la mayoría de los juicios se resuelven dentro de los 10 días.

También, se nota que en este mes los juicios les toman como máximo, para su resolución, 50 días tal como se muestra en el histograma.

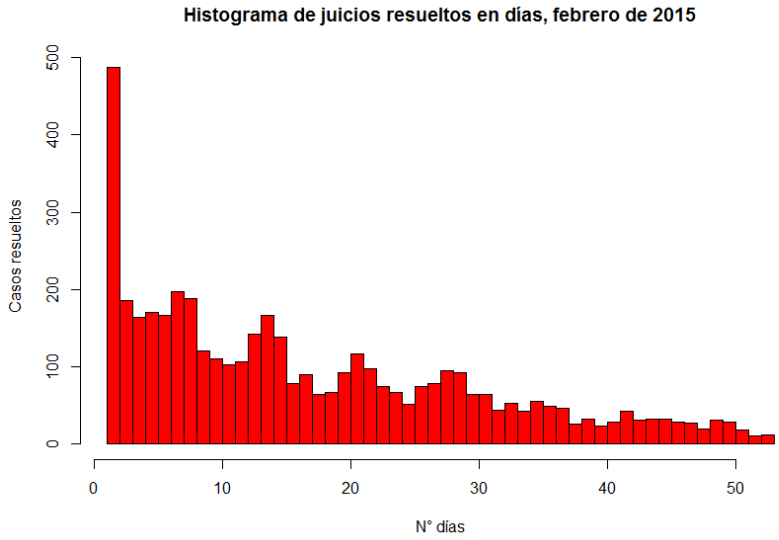

Figura 11. Histograma de juicios resueltos en días, febrero de 2015.

Al observar el histograma del mes de marzo de 2015, se observa que la mayoría de juicios se concentran entre 1 y 20 días para su resolución mientras que el tiempo máximo para resolución de juicios en este histograma es de 80 días lo que es aproximadamente tres veces mayor que enero y febrero.

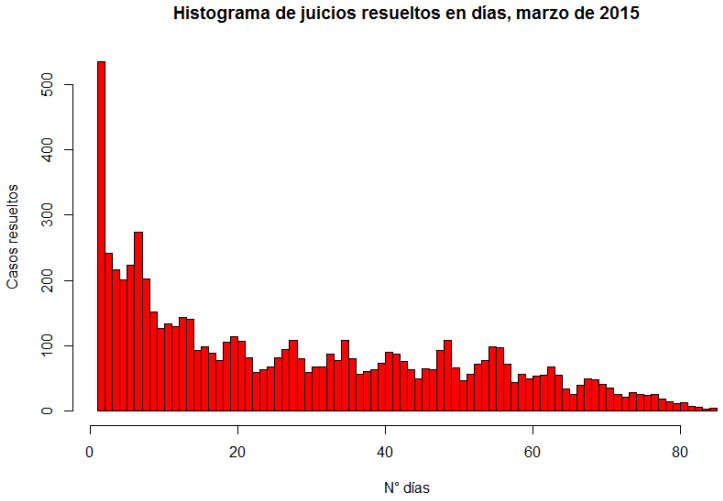

Figura 12. Histograma de juicios resueltos en días, marzo de 2015.

En este histograma de la Figura 13, se observa que los tiempos para resolución de causas se concentran entre 1 y 40 días. También se observa que hay tiempos que han llevado 120 días para resolver una causa. Esto implica que existieron juicios que iniciaron en enero y terminaron este mes.

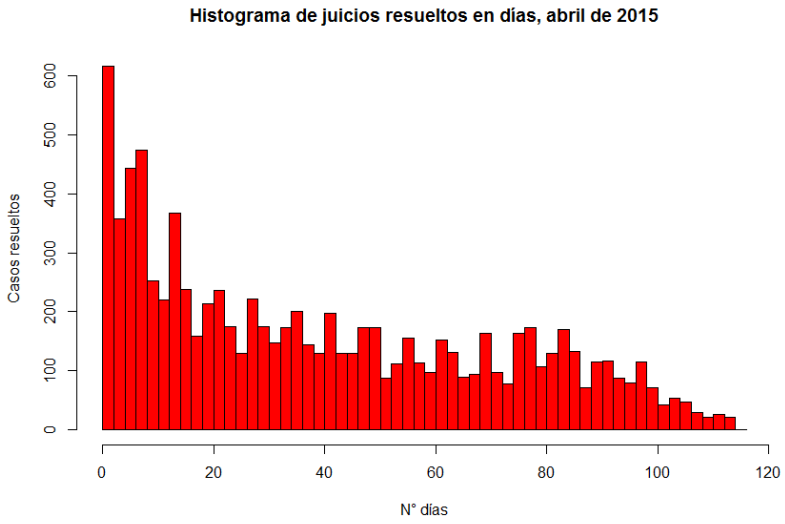

Figura 13. Histograma de juicios en días, abril de 2015.

En el histograma de mayo de 2015, el cual se representa por la Figura 14, se verifica que la concentración de juicios va entre 1 y 40 días para resolución de causas. Posiblemente son casos judiciales que iniciaron meses atrás y se resolvieron en este mes. Se observa que el tiempo máximo para este histograma fue de 140 días, lo cual indica que ha existido arrastre de juicios de meses anteriores.

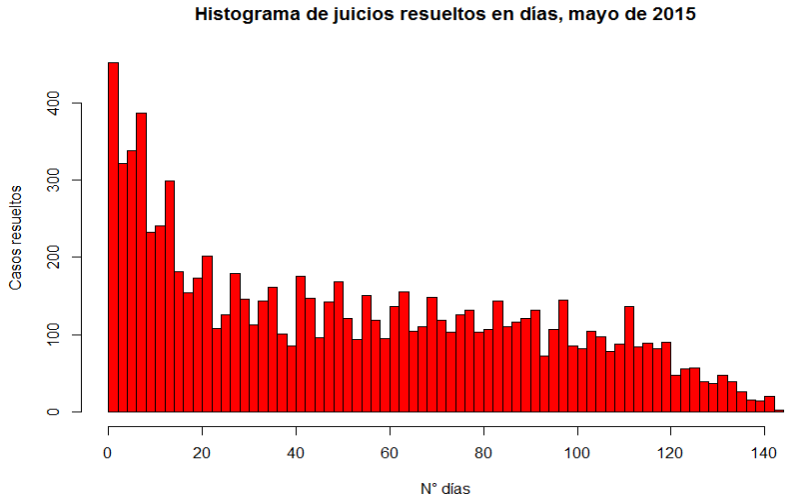

Figura 14. Histograma de juicios resueltos en días, mayo de 2015.

En la Figura 15, se tiene un histograma de junio de 2015. En este histograma se observa que los juicios resueltos se concentran en 50 días mientras que el periodo más largo que tuvo este mes es de 150 días.

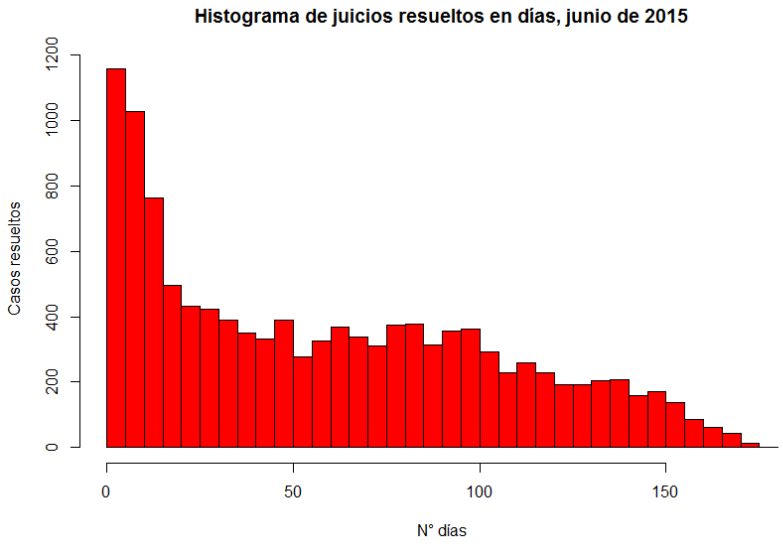

Figura 15. Histograma de juicios resueltos en días, junio de 2015.

En este histograma Figura 16, del mes de julio de 2015, se observa que hasta 1200 juicios se han resueltos en un tiempo menor a 50 días, mientras que luego presenta una normalización de 50 hasta 150 luego disminuye. Su tiempo máximo de resolución de causas para estos juicios es de 200 días.

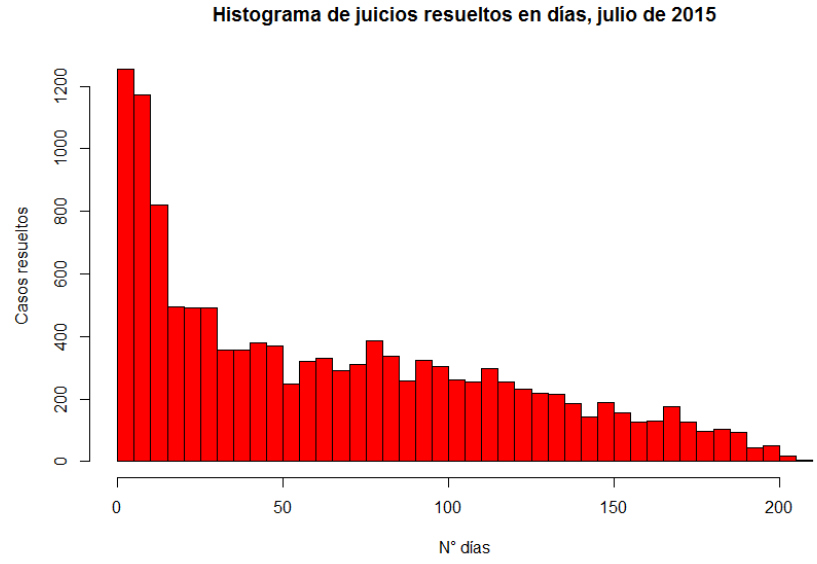

Figura 16. Histograma de juicios resueltos en días, julio de 2015.

En esta Figura 17, se observa que los juicios resueltos con mayor concentración van de 1 a 50 luego van disminuyendo hasta llegar a una cantidad máxima de 200 días para resolución de causas.

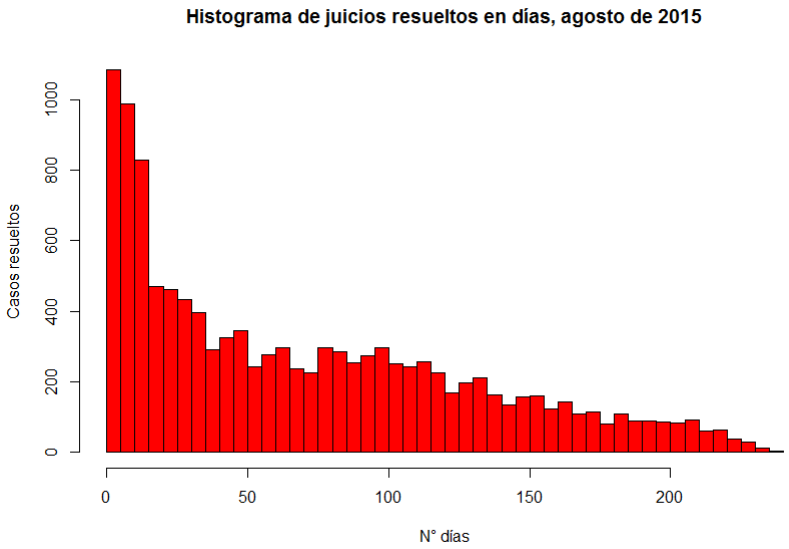

Figura 17. Histograma de juicios resueltos en días, agosto de 2015.

En el siguiente histograma de la Figura 18, del mes de septiembre se verifica que hay una gran concentración que oscila entre 1 y 100 días. Además, se observa que el tiempo de resolución máxima para los casos procesales es de 250 días.

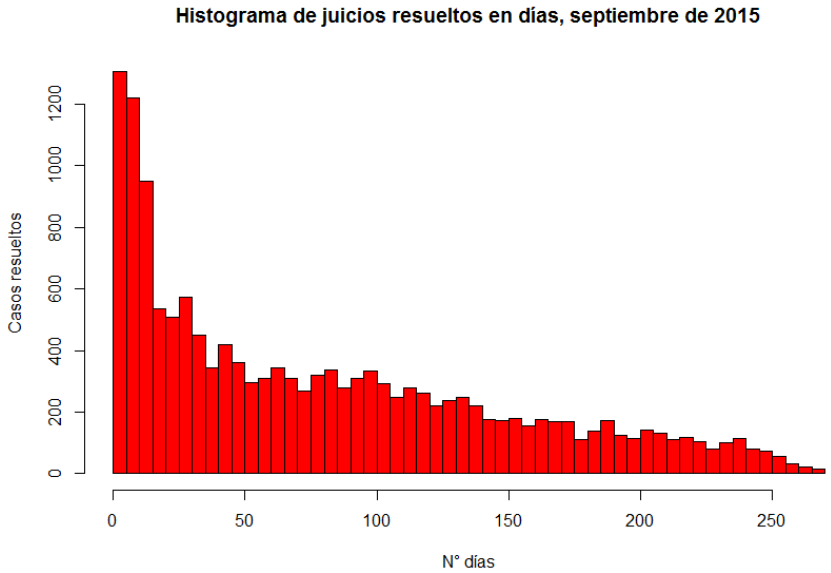

Figura 18. Histograma de juicios resueltos en días, septiembre de 2015.

En la Figura 19, se observa que gran concentración de juicios resueltos que oscilan de 1 a 100 días tal como el mes de septiembre. Luego sufre una disminución hasta llegar aproximadamente a 300 días para la resolución de causas.

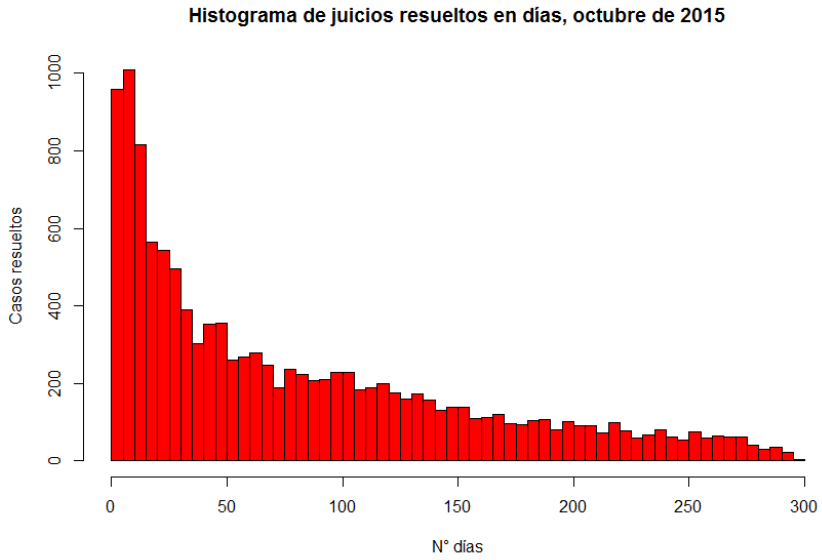

Figura 19. Histograma de juicios resueltos en días, octubre de 2015.

El siguiente histograma de la Figura 20, del mes de noviembre de 2015, se observa que la concentración de juicios que oscilan entre 1 y 100 días y sucede lo mismo para agosto, septiembre y octubre del mismo año. También, se observa que el periodo máximo para esta gráfica es de 300 días.

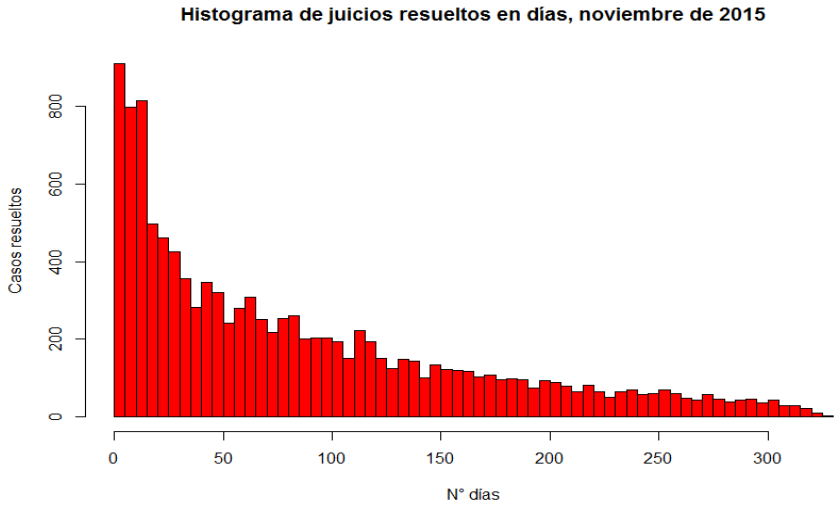

Figura 20. Histograma de juicios resueltos en días, noviembre de 2015.

Finalmente, el histograma de la Figura 21, de diciembre de 2015, el cual se observa que hay una gran cantidad de concentración de juicios resueltos que oscila desde 1 hasta 100 días luego se observa un descenso de juicios resueltos que llega como máximo hasta los 350 días.

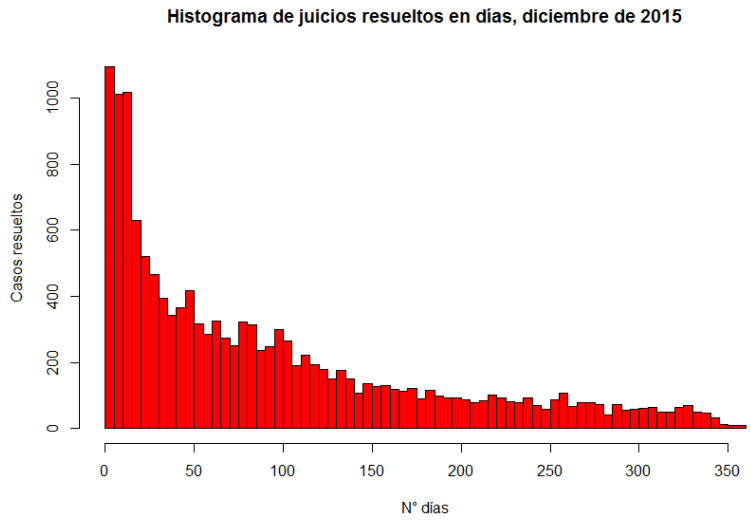

Figura 21. Histograma de juicios resueltos en días, diciembre de 2015.

Tendencias

A continuación, se realiza una exploración de tendencias de los datos por mes del año 2015. En la Figura 22, se observa un ascenso de los juicios resueltos en este mes. Se observa que el mes inicia resolviendo 2 causas (como promedio) pero conforme avanza el mes esto se evoluciona a 25.

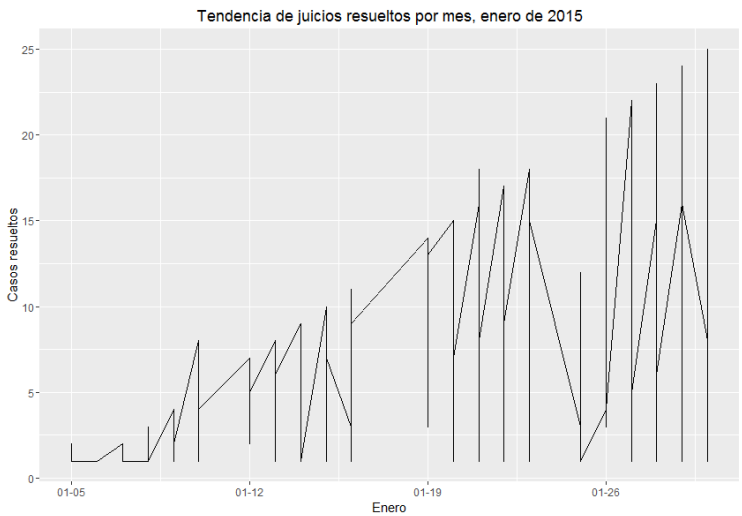

Figura 22. Tendencias de juicios resueltos por mes, enero de 2015.

En la Figura 23, se observa una tendencia de juicios resueltos que tiene como partida los 25 días y tiene un máximo de 50 días de causas resueltas durante el mes.

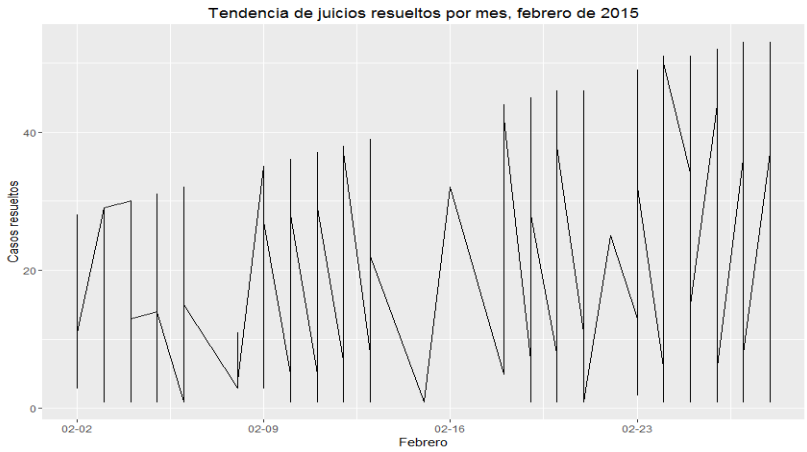

Figura 23. Tendencia de juicios resueltos por mes, febrero de 2015.

Al observar la Figura 24, de tendencia de juicios resueltos del mes de marzo de 2015, se verifica que los juicios resueltos tienen una amplitud mayor a la del mes de febrero. Eso quiere decir que resolvió más causas al iniciar el mes y al final termina con más 80 causas resueltas como promedio.

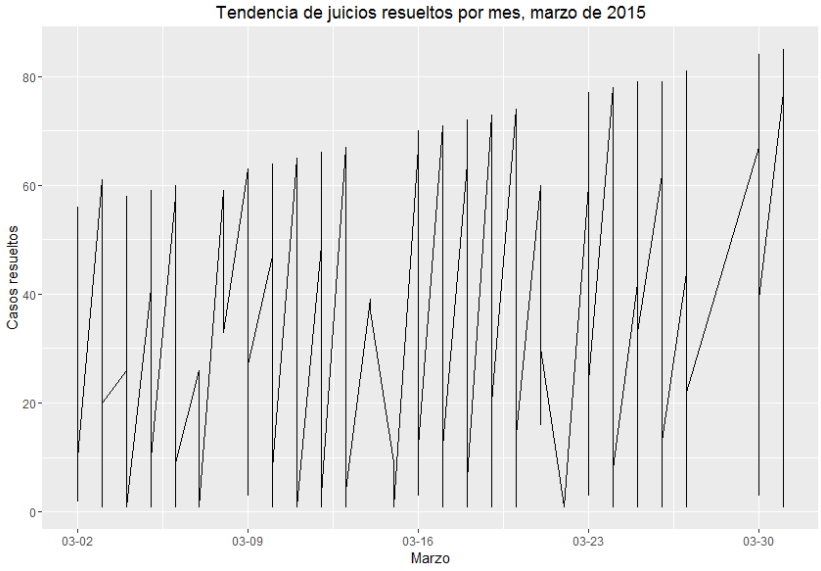

Figura 24. Tendencia de juicios resueltos por mes, marzo de 2015.

En la tendencia del mes de abril de la Figura 25, se observa que hay juicios que tienen como inicio de causas resueltas 90 y como fin llegan a las 120.

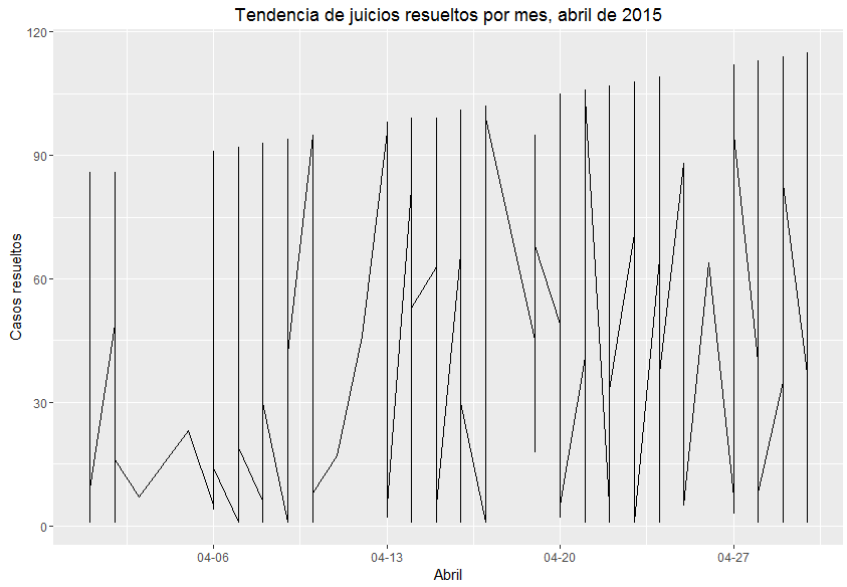

Figura 25. Tendencia de juicios resueltos por mes, abril de 2015.

En esta Figura 26, se observa las tendencias del mes de mayo, la cual presenta una gran concentración entre los días 11 y fin mes de mes. Aquí se nota que se inicia resolviendo 50 causas por día. Pero conforme se avanza en el mes, se verifica que llegan a 150 aproximadamente.

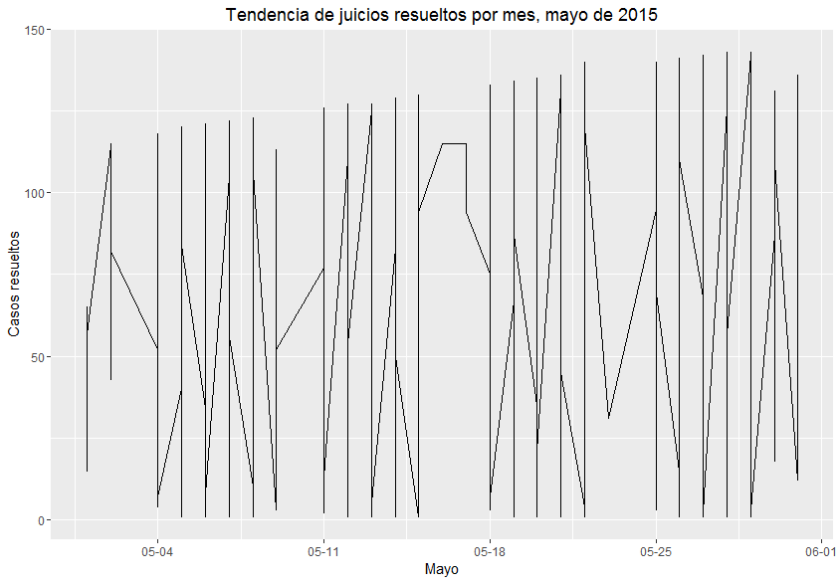

Figura 26 . Tendencia de juicios resueltos por mes, mayo de 2015.

En la Figura 27, se aprecia una concentración entre los días 1 y 8 del mismo mes y otra entre 15 y 29. También, se observan variaciones dentro de esas tendencias. En este mes, los promedios de las causas resueltas varían entre 50 y 150 como promedio.

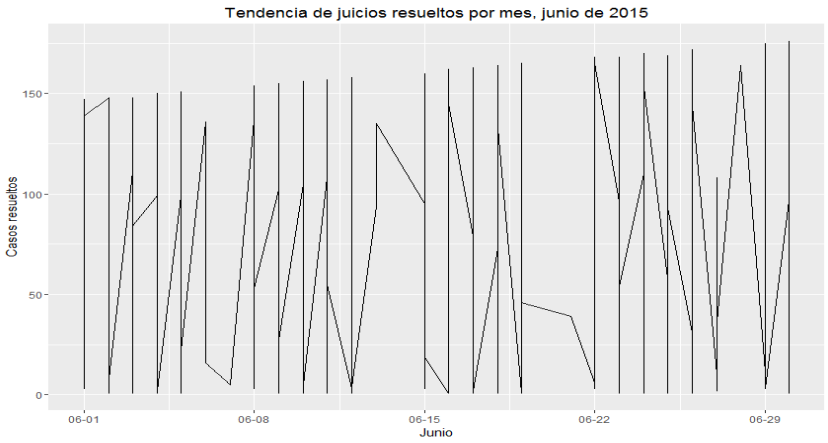

Figura 27. Tendencia de juicios resueltos por mes, junio de 2015.

Al apreciar la tendencia del mes de julio, en la Figura 28, se observa que existe una gran cantidad de juicios resueltos que como promedio van de 150 a 200 causas. Además, se observa que es un mes de mucha actividad. Este mes al igual que junio tienen un inicio de 150 causas resueltas por días.

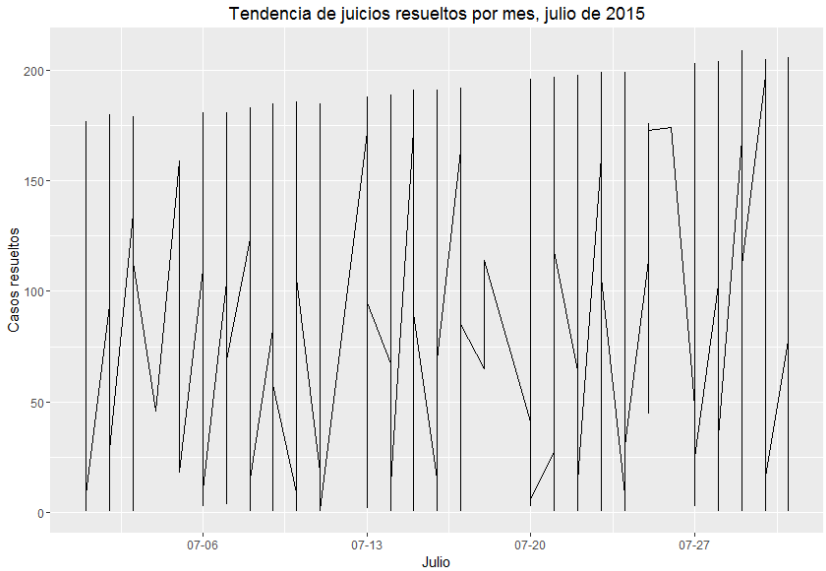

Figura 28. Tendencia de juicios resueltos por mes, julio de 2015.

En la Figura 29 del mes de agosto, se aprecia anomalías entre 8 y 10 de agosto, los cuales se debe a días feriados. También, se observa que los casos sobre pasan las 200 causas resueltas por días.

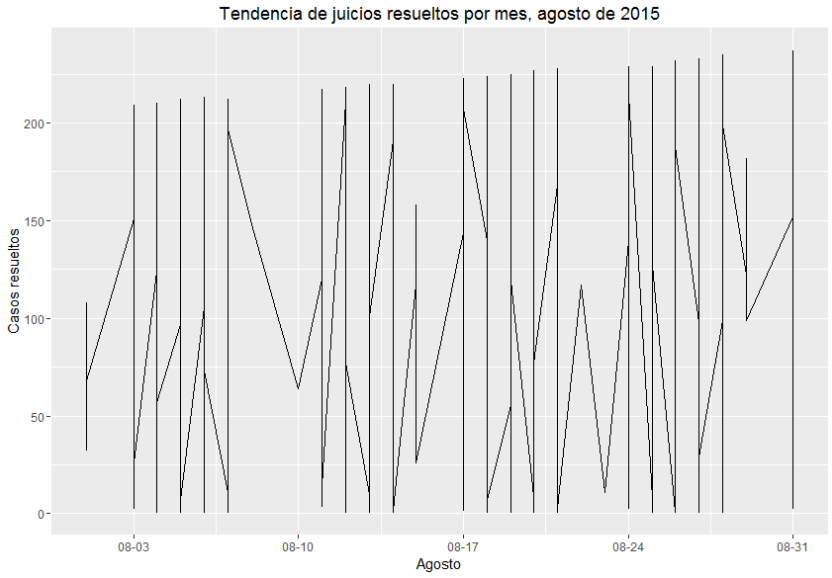

Figura 29. Tendencia de juicios resueltos por mes, agosto de 2015.

En la Figura 30, existe mucha actividad desde inicio a fin. Se observa que al inicio de mes se han resuelto aproximadamente más de 200 casos y así continua hasta fin de mes. Por eso se dice que a partir de julio son los meses que más juicios se resuelven.

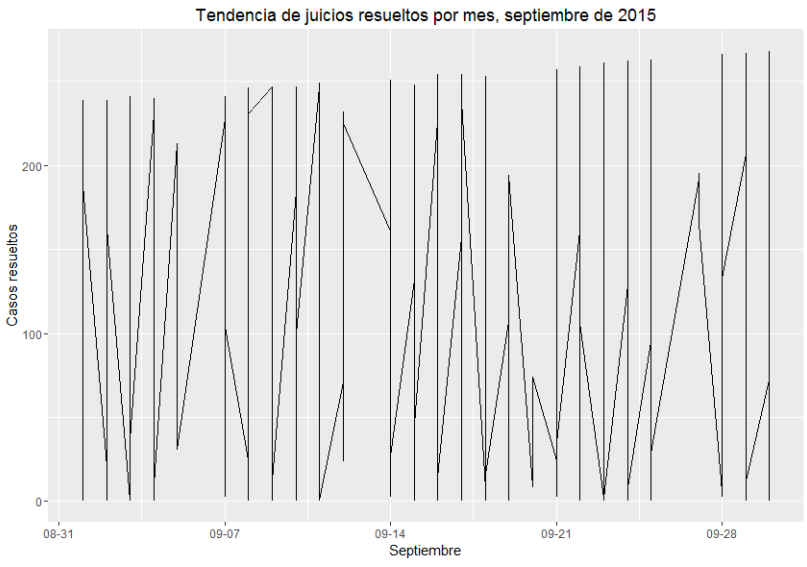

Figura 30. Tendencia de juicios resueltos por mes, septiembre de 2015.

En la Figura 31, se aprecia que el mes de octubre inicia resolviendo 200 causas judiciales, pero a finales de mes termina resolviendo 300 causas procesales como promedio.

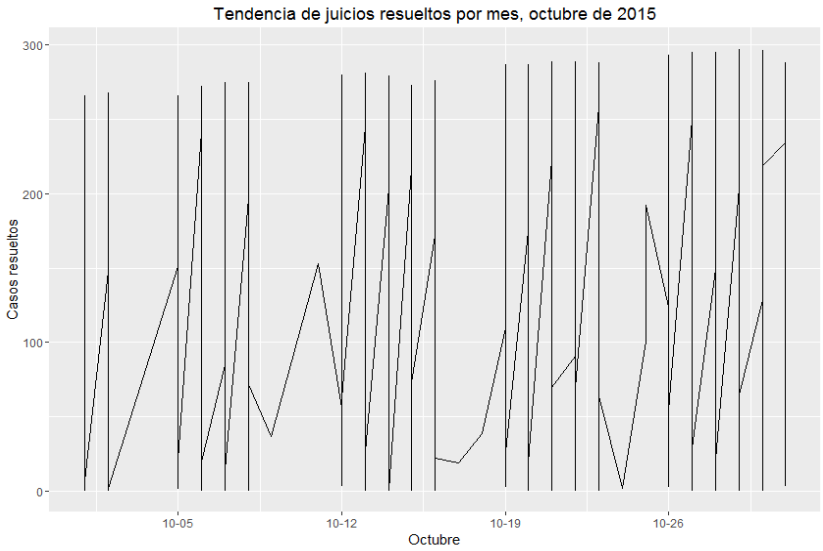

Figura 31. Tendencia de juicios resueltos por mes, octubre de 2015.

En el mes de noviembre de la Figura 32, se observa que se inició resolviendo 300 causas y luego bajaron a 100 pero luego subió superando las 300 causas superando a los meses anteriores. Además, se observa que los días con mayor actividad van desde 9 a 16 y 11 a 30 del mismo mes.

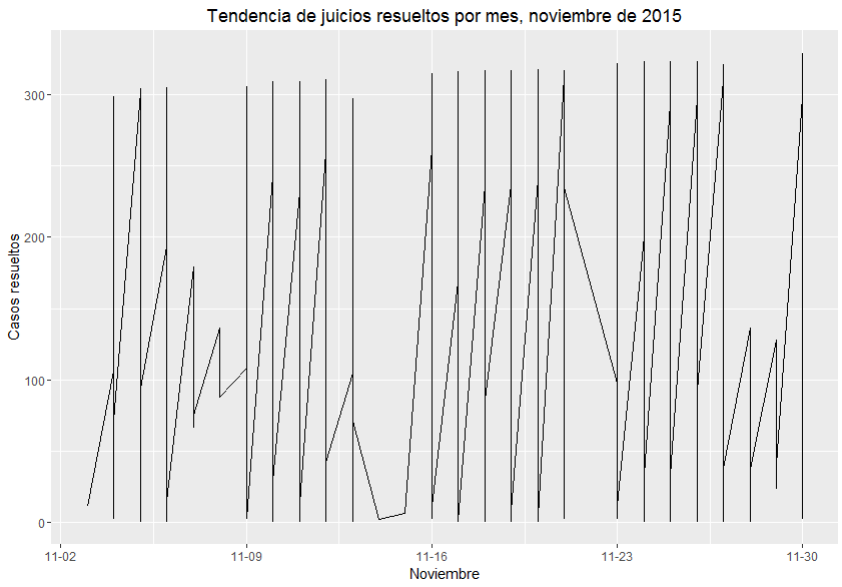

Figura 32. Tendencia de juicios resueltos por mes, noviembre de 2015.

Finalmente, en la Figura 33, al analizar el último mes del año 2015, se observa que existe mayor concentración que el mes de noviembre. El mes inicia resolviendo más de 300 causas superando a noviembre y a los meses anteriores. También, se observa que hay picos que posiblemente son causados por días feriados. Al final de mes, se verifica que la resolución de causas llega a 400.

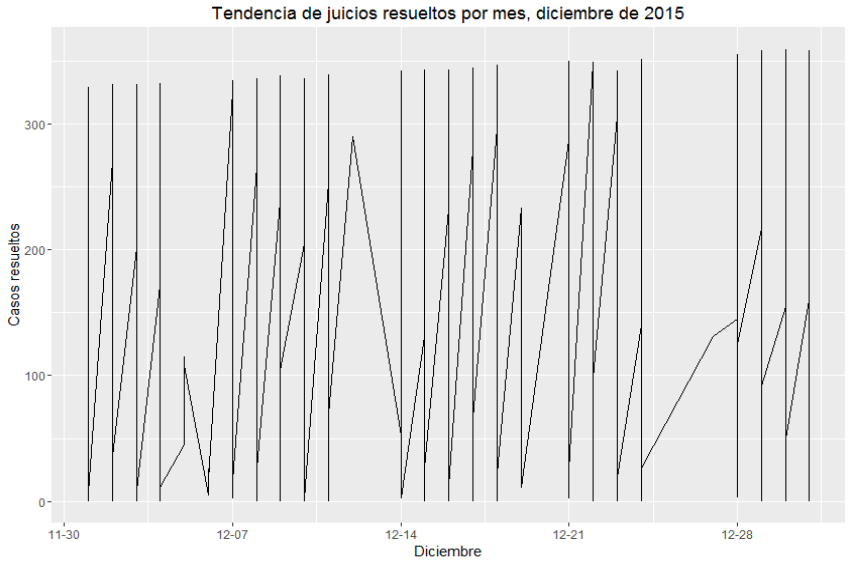

Figura 33. Tendencia de juicios resueltos por mes, diciembre de 2015.

Histograma del año 2015

La Figura 34 muestra la distribución normal de los procesos judiciales resueltos en la materia "Familia Niñez y adolescencia" de todo el año 2015. Se observa que existe una variación, pero son pequeñas con respecto al análisis de los histogramas mensuales del mismo año.

El Figura 34 explica que la mayor cantidad de juicios resueltos se concentra de 1 a 100 días y el resto como es de esperar son casos extraordinarios donde posiblemente la persona ha apelado al no estar conformes por lo que se produce una prolongación del tiempo de resolución de los juicios. (Ver Anexo 2. Literal C).

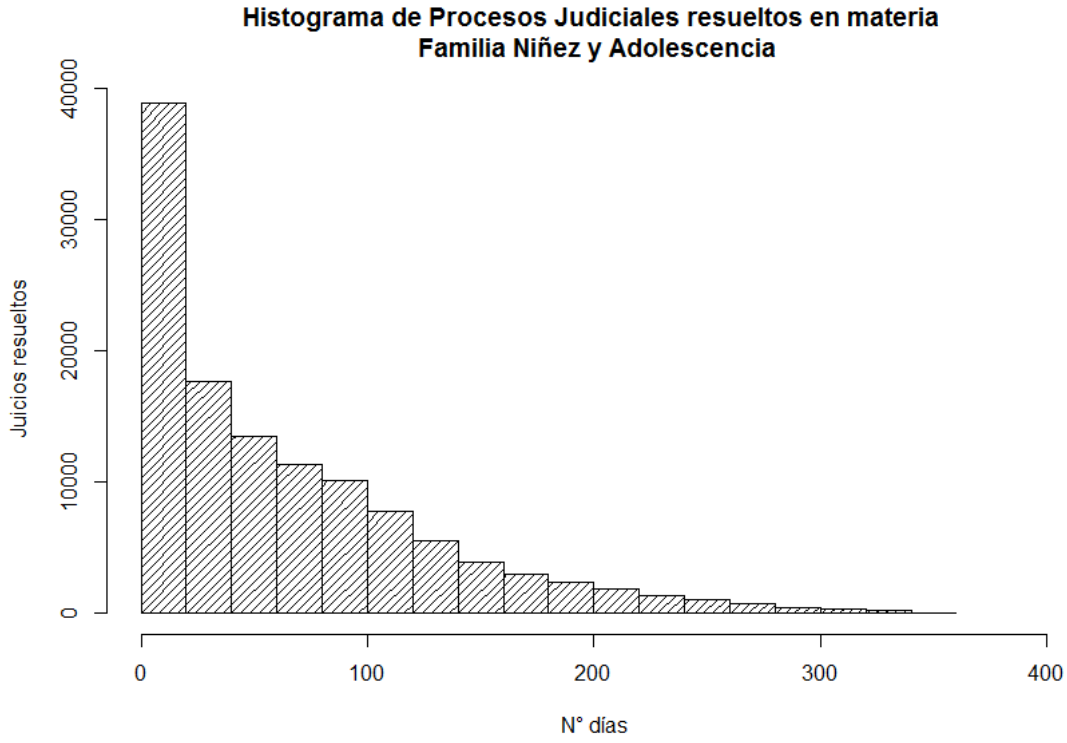

Figura 34. Histograma de procesos judiciales resueltos en materia Familia Niñez y Adolescencia.

En la Figura 35, se observa la tendencia de los juicios resueltos vs número de días en la materia "Familia Niñez y Adolescencia". Se identifica que la mayor concentración de juicios se encuentra entre los intervalos de 25 y 150 días. En otras palabras, la tendencia de resolución de juicios tiene una atendencia a bajar. (Ver Anexo 2, Literal D)

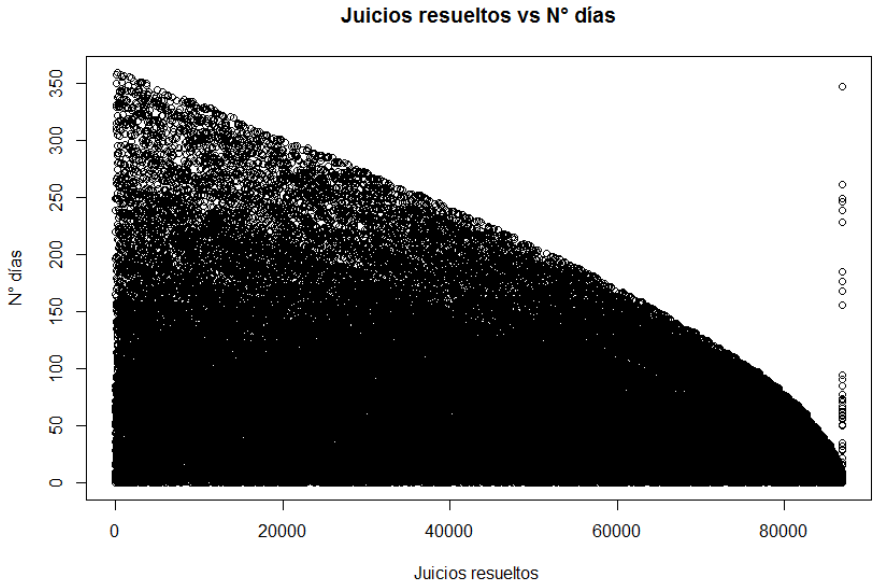

Figura 35. Tendencia de juicios resueltos.

En la Figura 36, se puede apreciar la curva de distribución normal de los juicios resueltos en la materia "Familia Niñez y Adolescencia" del año 2015. Esta curva se caracteriza por tener su centro en los intervalos de 70 a 90 días, lo cual da entender que es la mayoría de los juicios se resuelven entre esta cantidad de tiempo. (Ver Anexo 2, Literal E)

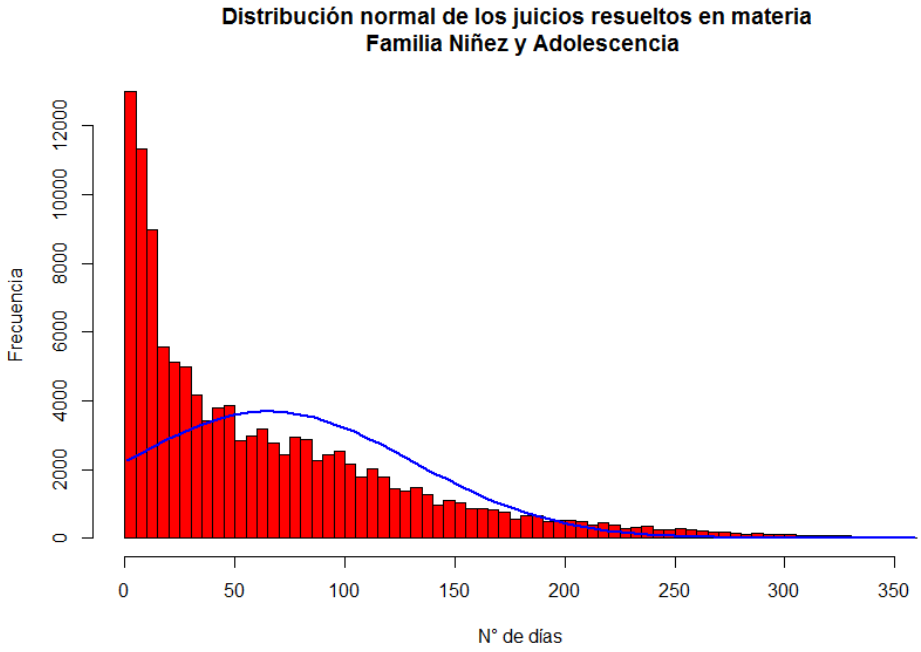

Figura 36. Distribución normal de los juicios resueltos.

En la Figura 37, muestra la cantidad de juicios resueltos, en la materia “Familia Niñez y Adolescencia” durante los meses del 2015. Se observa claramente, que, a inicios del 2015, la resolución de procesos judiciales fue muy baja pero conforme los meses avanzan, fue creciendo de forma lineal las resoluciones de los juicios. También se puede apreciar que el mes con más juicios en este mes de septiembre con un total de 10.640 casos a nivel nacional. Esta grafica explica el porqué de los tiempos bajos en la gráfica de la Figura 10, 11, 12, 13 y 14.

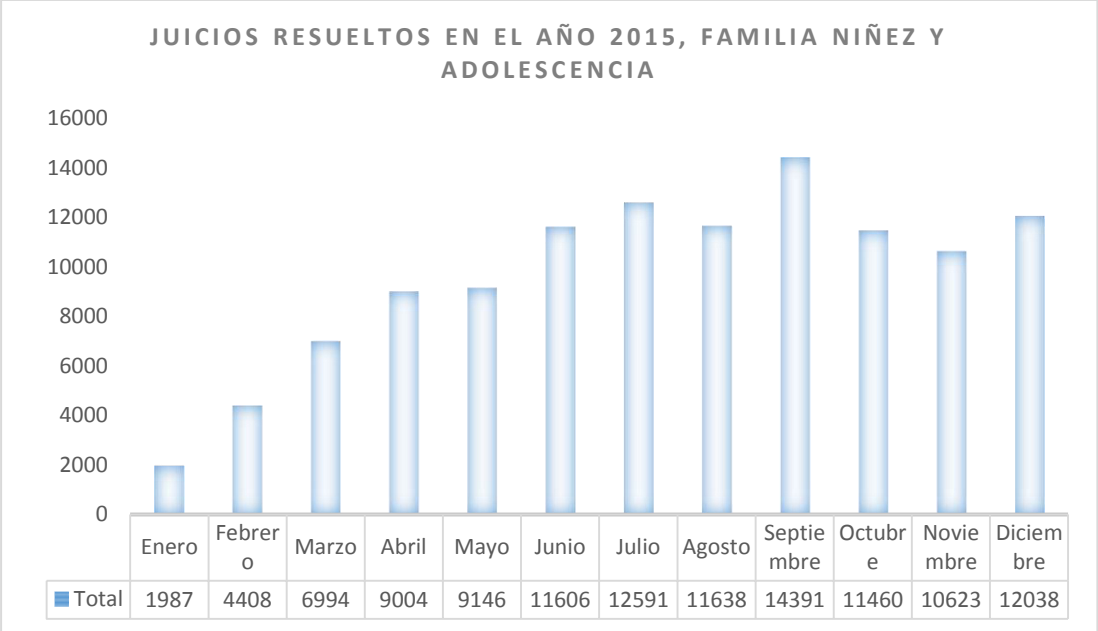

Figura 37. Juicios resueltos por mes en la materia "Familia Niñez y Adolescencia".

3.3.15 Verificar la calidad de los datos

En el transcurso del análisis de los registros se observó que existían total de días igual a 0, es decir que resolver un juicio no llevo ni un día. Además, existían fechas de resoluciones del 2016. Pero muy a parte de esos datos encontrados se evidenciaron datos “outliers” a través de diagrama de cajas tal como se observa en la Figura 38.

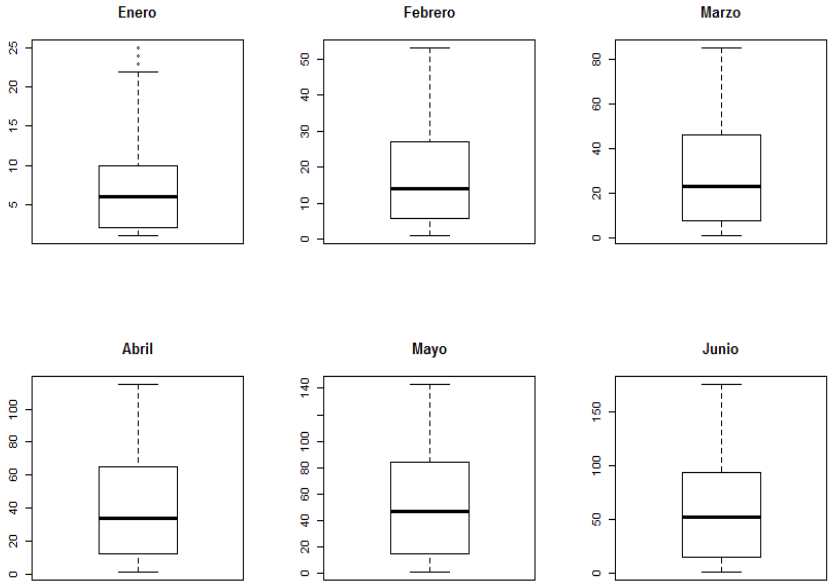

Figura 38. Diagrama de Cajas Enero-Junio año 2015, materia familia niñez y adolescencia.

En la gráfica, se observa datos atípicos en el mes de enero, pero el resto de meses que van desde febrero hasta junio no se observa. Hay que mencionan que de acuerdo a Turkey 1969, cualquier valor fuera del cuarto cuartil es considerado dato a típico. (Ver Anexo 3, Literal A).

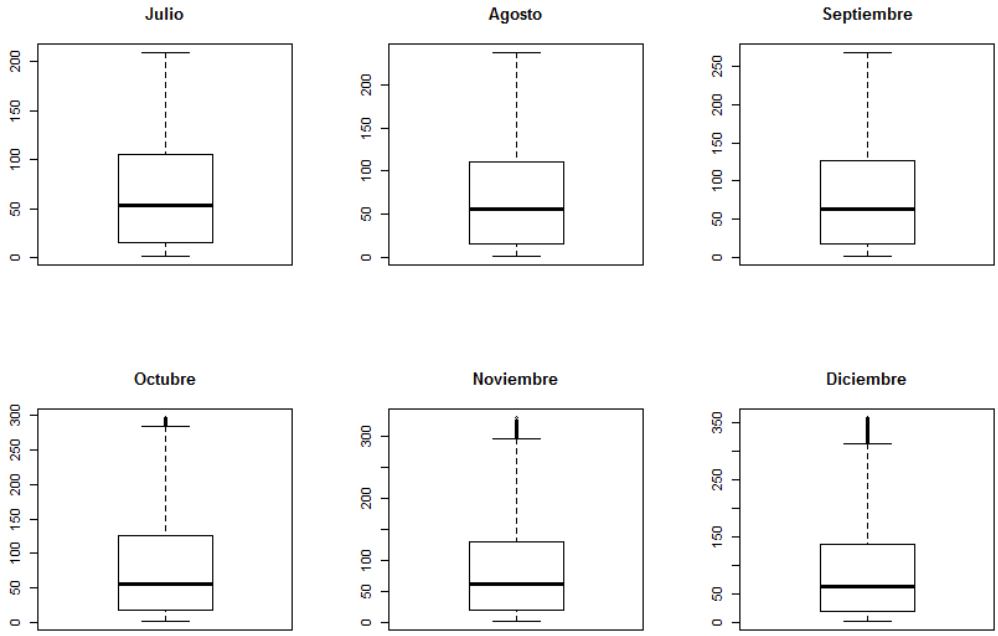

Figura 39. Diagrama de cajas de Julio-Diciembre año 2015, materia familia niñez y adolescencia.

En cambio, en esta Figura 39, se aprecia datos atípicos que van desde el mes de octubre hasta diciembre del año 2015.Se recuerda que estos datos "outliers" al ser analizados y representados a través de modelos darán un modelo muy alejado a la realidad. Entonces, de acuerdo a las observaciones realizadas en estos diagramas de caja, se procedió a verificar el porcentaje de representación de registros "outliers" sobre el total de los datos. En la Figura 40, se observa que existen un 98% de datos limpios. Es decir, registros no atípicos, los cuales son útiles para el análisis. Pero el 2% de la figura indica que están compuestos por datos atípicos.

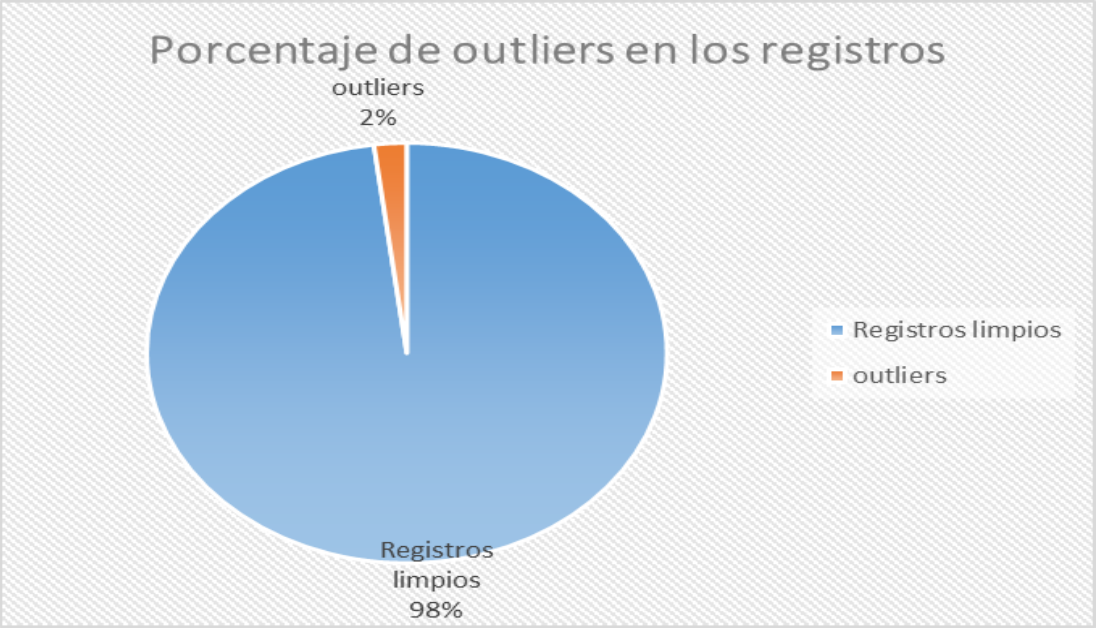

Figura 40. Porcentaje de datos outliers.

Por lo tanto, en base a las observaciones realizadas en las gráficas se procedió a no tomar en cuenta a los datos atípicos por ser un porcentaje muy pequeño y otra porque estos resultados alejan de la realidad al modelo encontrado. Para encontrar que datos eran atípicos, o desde que rango, se observó las grafica de la Figura 38 y 39, y se determinó que los datos mayores a 250 no deberían ser considerados para el análisis porque ellos representan el 2% del total de los registros. Para comprobar la calidad de los datos para realizar el análisis se procedió a realizar un nuevo diagrama de caja, pero esta vez ignorando a los datos atípicos y se obtuvo los siguientes resultados. (Ver Anexo 3, Literal B)

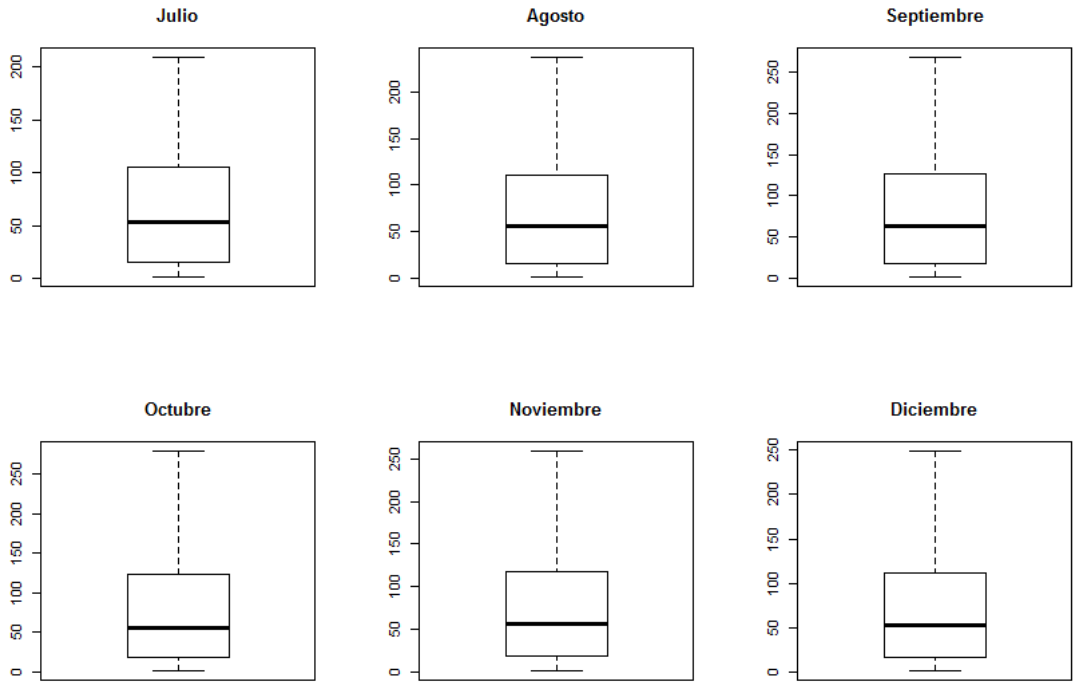

Figura 41. Corrección datos outlier.

En la Figura 41, se observa que los datos outlies desaparecieron y esto significa que los datos se encuentran listo para arrojar resultados cercanos a la realidad.

3.3.16 Preparación de datos

De acuerdo a CRISP-DM, se debe concentrar en la selección de los datos, los cuales serán utilizados en los modelos. Luego se debe realizar limpieza, construcción si es necesario. Y finalmente se los debe integrar y formatear. Hay que mencionar que, para este caso, algunas de estas tareas se pasaran por alto debido a que la información está completa.

3.3.17 Selección de los datos

En esta parte se debe seleccionar los registros con lo cual trabajarán los modelos, para ello se ocupará el 100% de los registros de la Tabla "proceso_judicial". A continuación, se presentan los atributos empleados para el análisis de minería:

Tabla proceso_judicial

- FechaResolucion

- NumDias
- IdTipoAccion
- IdMateria

Como podrán observar no todos los campos fueron tomados en cuenta para trabajar con los modelos porque todo va en función de los objetivos de la minería de datos que se definieron en la sección de la comprensión del negocio de la metodología.

3.3.18 Limpieza de los registros

Debido a que los registros con los que se dispone para trabajar con los modelos esta limpios, se considera que no es necesario una limpieza. En partes anteriores se mencionó que existían número de días con igual a cero y fechas fuera del intervalo del 2015, pues estos tan solo se los ignora a la hora de realizar la consulta para el análisis con los modelos, además no aportan nada relevante a la investigación.

A continuación, se tiene la Figura 42, donde explica gráficamente el filtro que se debe realizar a la información para su análisis. En el Anexo 4, se encuentra el script que hace el paso de la Tabla "proceso_judicial" a "analisis_judicial" permitiendo tener solo aquellos procesos judiciales que se encuentren dentro del año 2015 cuyo tiempo de proceso sea diferente a cero, menor e igual a 250 y sean de la materia familia niñez y adolescencia.

Se debe aclarar que no se requirió de herramientas especializadas para la migración tales como Pentaho, Oracle Datamining, Sql Server Analysis entre otros; porque el migrado de datos de una a la otra Tabla no requiere tantas cosas, más que una sentencia de llenado con los parámetros de selección especificados en el Anexo 4.

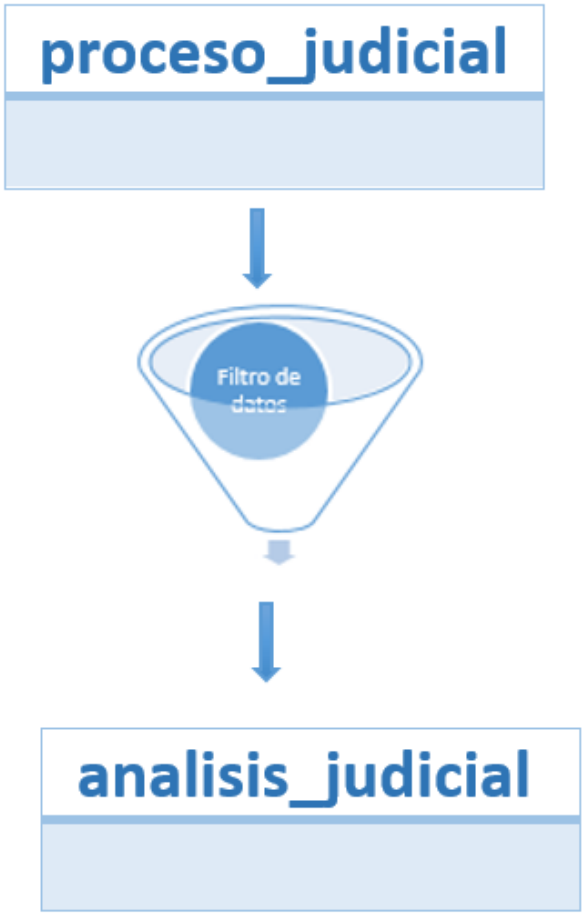

Figura 42. Limpieza de datos.

3.3.19 Construir datos

Atributos derivados

Debido a que los datos a ser analizados se encuentran con su respectivo tipo de datos, no se realizó ningún tipo de transformación ni tampoco se agregaron campos adicionales para cálculos de fechas.

Registros generados

Para esta fase, la información con la que se dispone es necesaria y cubre la necesidad, por lo tanto, no se ha visto la necesidad de crear nuevos atributos.

3.3.20 Integrar datos

Se creó la estructura “analisis_judicial” donde se almaceno los registros después de ser limpiados. En la Tabla 3, se observa los atributos de la Tabla creada donde están almacenados los registros después del proceso de limpieza.

Tabla 3
Análisis judicial.

CAMPO	TIPO DATO
IdMateria	Numérico
IdTipoAccion	Numérico
NumDia	Numérico
FechaResolucion	Fecha

3.3.21 Formatear datos

Debido cuando se realizó el paso del archivo de Excel a nuestra base se realizaron las transformaciones necesarias no se vio la necesidad de formatear algún campo adicional. Por lo que no es necesario realizar esta actividad.

3.4 Modelado

En esta fase se seleccionará la técnica, se creará el script para los modelos, se trabajará con los datos y se ejecutará los modelos para ver su comportamiento. Además, se evaluarán los resultados desde el punto de vista de la minería.

3.4.1 Selección de la técnica de modelado

En esta investigación, se han seleccionado dos técnicas de modelado con el fin de comparar los resultados y seleccionar la mejor. En la vida real es así, porque no solo basta con una porque puede que haya mejores ajustes con otras técnicas. Para este propósito los modelos seleccionados son: LM y SVM.

3.4.2 Generación de la prueba para los modelos obtenidos

Para probar la calidad de los modelos obtenidos se va a utilizar primero la técnica de error absoluto medio (MAE) que consiste en una sumatoria de la diferencia del modelo menos el valor real para posteriormente dividirse para el total de predicciones. La Figura 43 a continuación expresa lo que literalmente se explicó.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Figura 43. Fórmula de Error Absoluto Medio

Fuente: ***(Montserrat, 2001)*.**

Otra de las técnicas a ser empleada es la de error cuadrático medio (RMSE) que de acuerdo a los estadísticos es de mucha utilidad para ver las variaciones de error de un modelo dado. Esta consiste en la raíz cuadra de la sumatoria de las diferencias del valor predicho por el modelo menos el valor real, todo eso elevado al cuadrado, luego se divide por el número de predicciones realizadas. A continuación, en la Figura 44, se expresa la fórmula del Error Cuadrático Medio.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Figura 44. Fórmula del Error Cuadrático Medio

Fuente: ***(Montserrat, 2001)*.**

R ofrece al usuario la opción de leer archivos csv donde se tiene la opción de determinar la cantidad de registros a ser leídos ya que los archivos son modificables directamente por el usuario y es quien decide el porcentaje de registros para entrenamiento y prueba. CRISP-DM no tiene establecido o claro el porcentaje de

datos que se utilizaran para prueba y entrenamiento, pero los expertos sugieren que se utilice el 40% para las pruebas reales y el resto para entrenamiento, pero en este caso se utilizará todos los datos para pruebas y entrenamiento debido a que la cantidad de registros a analizar no es muy extensa.

3.4.3 Construcción del modelo

Según la metodología CRISP-DM, en este apartado se debe concentrar en la generación del modelo y la herramienta para llevar a cabo la ejecución de ellos. En este caso, como ya se había mencionado antes, se ha escogido la herramienta R porque se está familiarizado con el entorno y es una herramienta muy flexible y tiene una gran cantidad paquetes de minería de datos que ahorrarán mucho tiempo.

Ajustes de Parámetros

En esta parte se va a desarrollar los modelos para cada uno de los objetivos definidos en la minería de datos y se irá resolviéndolos paso a paso.

- **Objetivo** 1. Verificar la existencia de la relación entre los propios registros de los procesos judiciales del año 2015 en materia de familia niñez y adolescencia.

1. Verificación de relación de casos resueltos entre los distintos tipos de acción en la materia familia niñez y adolescencia.

Al observar los datos de cada tipo de acción de la materia familia niñez y adolescencia se encontró una relación entre los tipos de acción: contencioso general, deprecatorio, especial, ordinario y verbal sumario. A continuación, se muestra la Figura 45 donde se evidencia los tipos de acción antes mencionados que comparten una tendencia muy similar. Si se observa detenidamente, se identifica que son los tipos de acciones con mayor demanda y tienen un tope de 250 como máximo para la resolución de las causas judiciales. (Ver Anexo 6, Literal A para la generación de las gráficas).

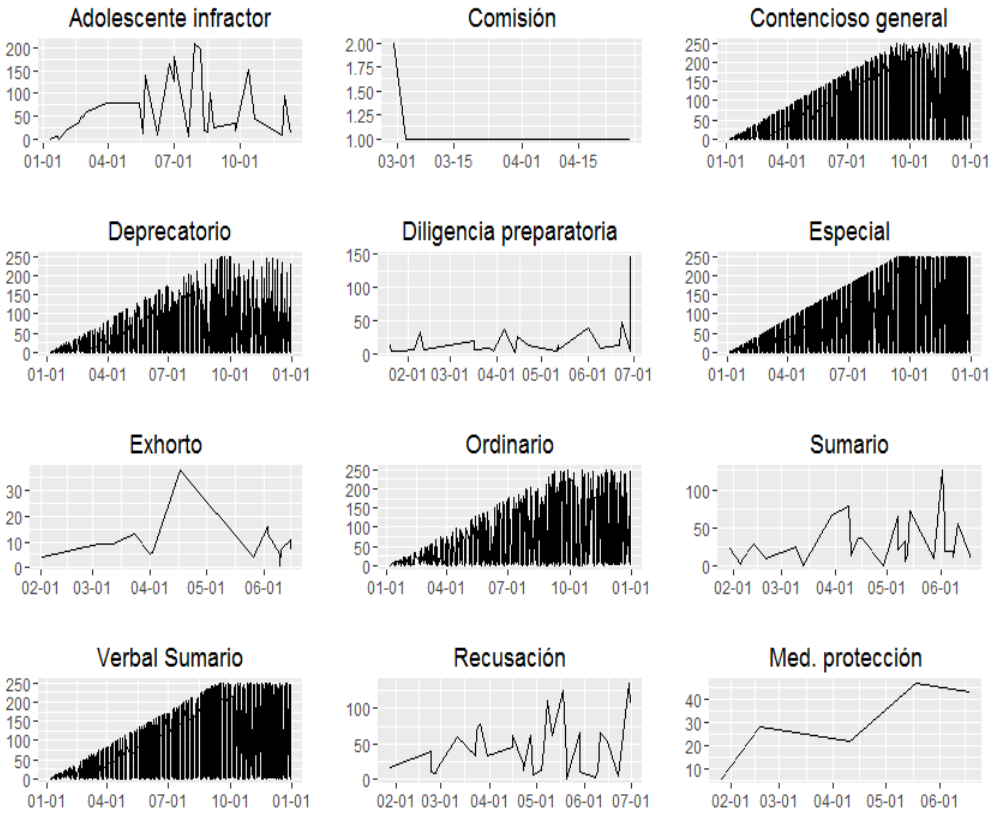

Figura 45. Tipos de Acción durante el año 2015.

Para plasmar la relación existente se procederá a realizar una regresión lineal para los tipos de acciones que comparten estas similitudes en su tendencia.

Regresión con el modelo LM

A continuación, se muestra la Figura 46, se puede notar los puntos negros que son los datos reales de los tipos de acción y los de color rojo son línea de tendencia después de aplicar el modelo lineal de regresión. Y las x de color azul son los puntos generados con el modelo. (Ver Anexo 6, Literal B. Script de la gráfica).

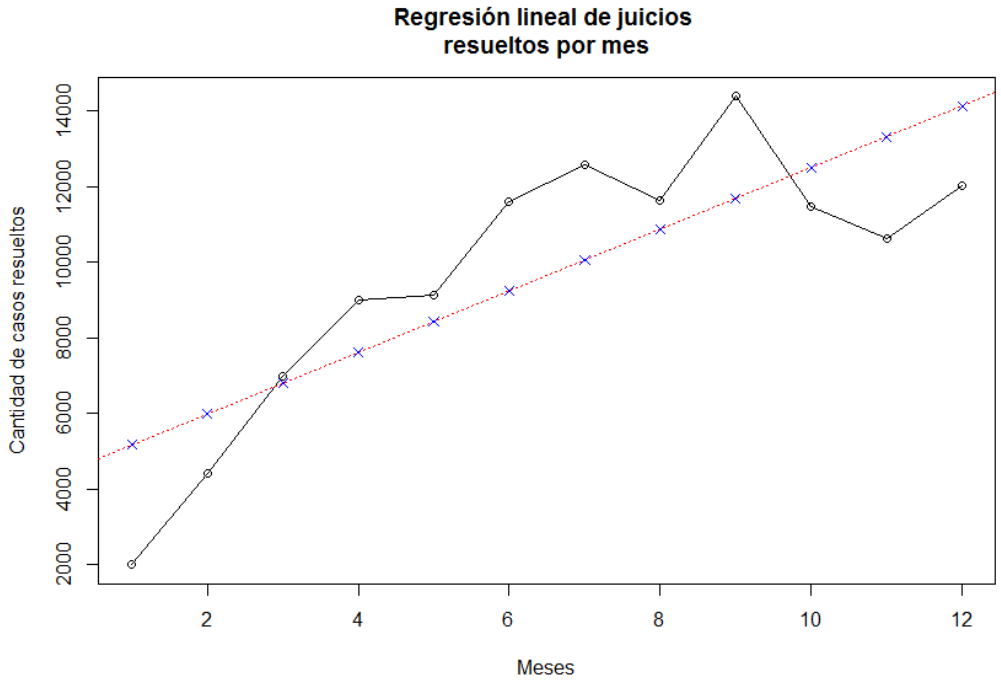

Figura 46. Regresión lineal de los tipos de acción relacionados.

El modelo ejecutado entregó la siguiente fórmula de la Figura 47(Ver Anexo 6, Literal B):

$$Y = 4358.2 + 815.2X$$

Figura 47. Ecuación lineal de los tipos de acción relacionados.

Donde:

Y: representa la cantidad de juicios resueltos.

X: los meses.

Tabla 4
Resumen del modelo de regresión lineal.

DATOS	
Intercepción	4358.2
Variable	815.2
RMSE	1994.134
MAE	1770.519

Regresión con el algoritmo SVM

Al probar los mismos datos con otro algoritmo de regresión como el de máquinas de vector de soporte conocido como SVM, se obtuvo el siguiente resultado. (Ver Anexo 6, Literal C. Script de la gráfica).

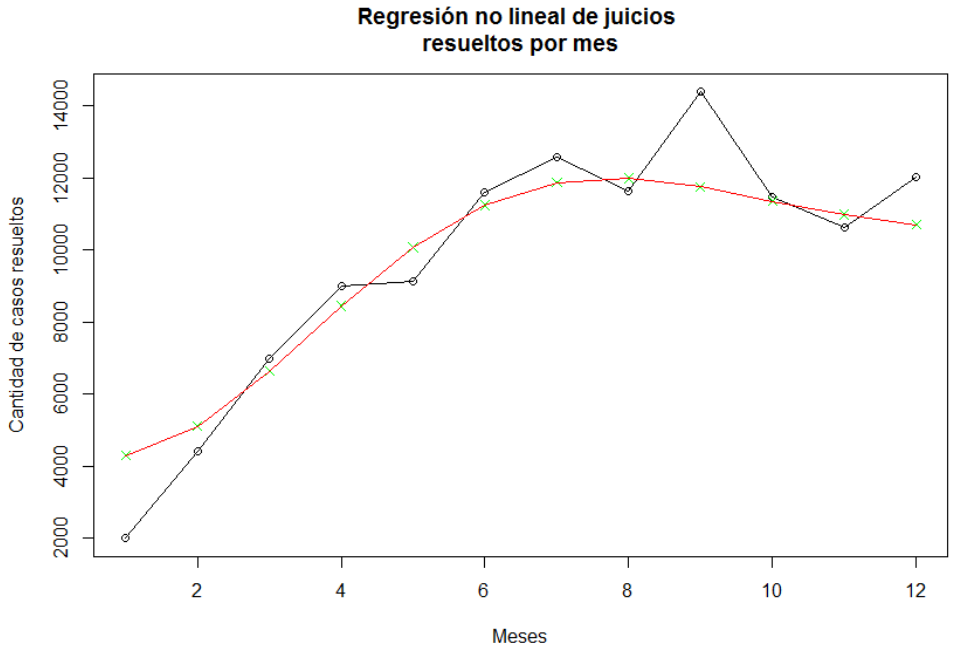

Figura 48. Regresión aplicando SVM a la gráfica de cantidad de juicios resueltos por mes.

Al observar la gráfica de la Figura 46, se nota la diferencia con la gráfica de la Figura 48. Esta última se ajusta más a los datos reales que la lineal. A continuación la Tabla de errores del modelo SVM:

Tabla 5
Resumen del modelo SVM.

DATOS	
Intercepción	NA
Variable	NA
RMSE	1180.685
MAE	892.4197

Al observar la Tabla 5, la sumatoria de errores tanto en RMSE y MAE es aproximadamente la mitad de la Tabla resumen del modelo lineal. La evaluación de estos modelos se dejará para la sección de "evaluación del modelo" donde se analizará la diferencia entre ellos y se escogerá el mejor.

2. Relación de los tiempos promedio de los tipos de acciones relacionados

Al observar la Figura 45, se observó los tipos de acciones con mayor concentración, y se verifico que estos parten de aproximadamente de 9 a 20 días para resolver un caso y esto se expande conforme avanza el tiempo, por tal motivo se procedió a generalizar a través de una ecuación que relacione el tiempo como la variable independiente para determinar el tiempo promedio de culminación de un juicio en los tipos de acciones observadas en la Figura 45.

Regresión con el modelo LM

A continuación, se observa la Figura 48 para analizar su tendencia junto con la regresión. (Ver Anexo 6, Literal D. Script de la gráfica).

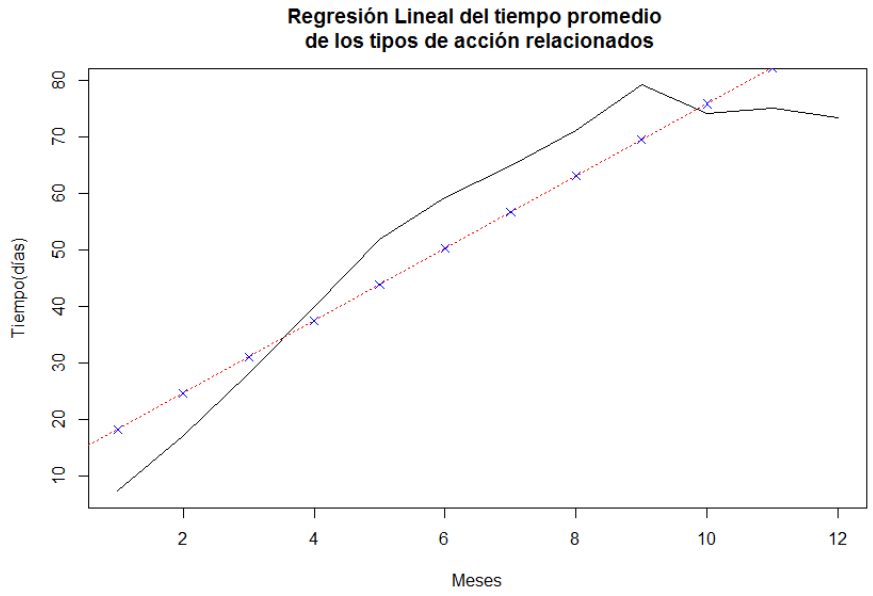

Figura 49. Regresión lineal del tiempo promedio en meses de los tipos de acciones relacionados.

Al observar la Figura 49, se visualiza que existen dos líneas. La negra son los datos reales y la roja es la regresión lineal generada por el modelo. Con respecto a la línea negra, se visualiza que existen juicios que pueden llegar a resolverse como máximo en 80 días, pero si se compara con la línea roja, generada por el modelo de regresión, se ve que se dispara, es decir que para el mes 12 los tiempos será mucho mayores que 80 días.

A continuación, la ecuación del modelo lineal:

$$Y = 11.86 + 6.4X$$

Figura 50. Lineal del tiempo promedio en meses de los tipos de acciones relacionados.

Donde:

Y: número de días en que termina un juicio

X: los meses.

Tabla 6
Resumen de la regresión lineal del promedio de atención en días de los tipos de acción.

DATOS	
Intercepción	11.8552
Variable	6.3993
RMSE	8.422817
MAE	7.5875

A simple vista, se puede decir que para análisis de un juicio toma un tiempo mínimo de 12 días y básicamente que a medida que avanza el tiempo se incremente el tiempo en resolver el juicio tal como lo describe la ecuación de esta sección.

Regresión con el algoritmo SVM

Ahora se prueban los datos con el algoritmo de regresión no lineal SVM para ver su tendencia con respecto a los puntos del conjunto de datos reales. A continuación, al observar la gráfica de la Figura 51, aplicando el modelo SVM de regresión no lineal. (Ver Anexo 6, Literal E. Script de la gráfica).

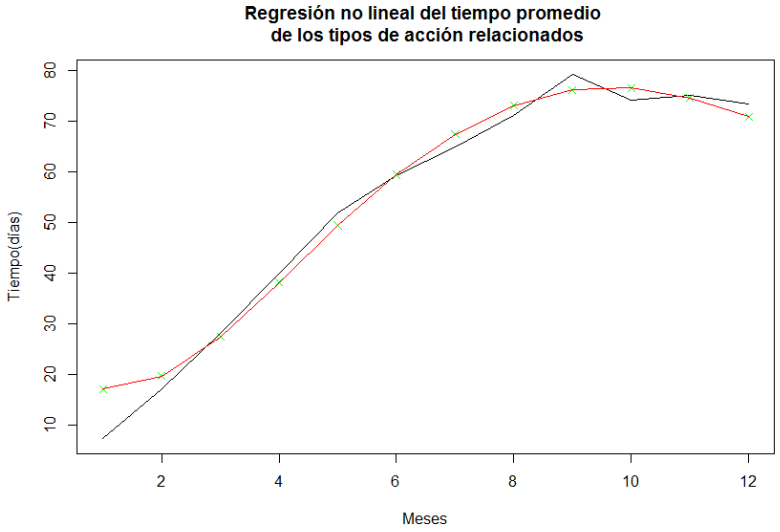

Figura 51. Regresión no lineal con algoritmo SVM.

En la gráfica de la Figura 51, se observa que el modelo de regresión no lineal SVM se aproxima a la curva de los datos reales. La diferencia es notable con respecto al modelo de regresión lineal. A diferencia de la lineal esta no pasa el tope de los 80 días para resolver un caso tal como la gráfica de frecuencias que se realizó en la exploración de los datos y tiene mucho sentido. A continuación, la Tabla resumen donde se describe, en este caso los errores RMSE y MAE.

Tabla 7
Resumen del modelo SVM.

DATOS	
Intercepción	NA
Variable	NA
RMSE	3.46

MAE	2.53

Los resultados de la Tabla 7, tienen un RMSE y MAE aproximadamente la mitad del modelo lineal lo cual es bueno. Entre menor se tenga un RMSE y MAE mucho mejor para el modelo de predicción.

- **Objetivo 2**. Formular un modelo predictivo de la duración promedio de un juicio a través del análisis de patrones de los procesos judiciales del año 2015 en materia de familia niñez y adolescencia.

En este objetivo se trata de encontrar un modelo predictivo para la materia familia niñez y adolescencia. Con los datos que se cuenta se tiene la relación entre los meses vs el tiempo promedio expresado en días de los juicios que puede llegar a tardar en resolver un juicio. En el Anexo 7, Literal A, se tiene el script realizado en PostgreSQL para obtener la información a ser analizada en la herramienta R.

Regresión con el modelo LM

A continuación, se presenta la gráfica con los datos reales vs la regresión. (Ver Anexo 7, Literal B. Script para generación de la gráfica).

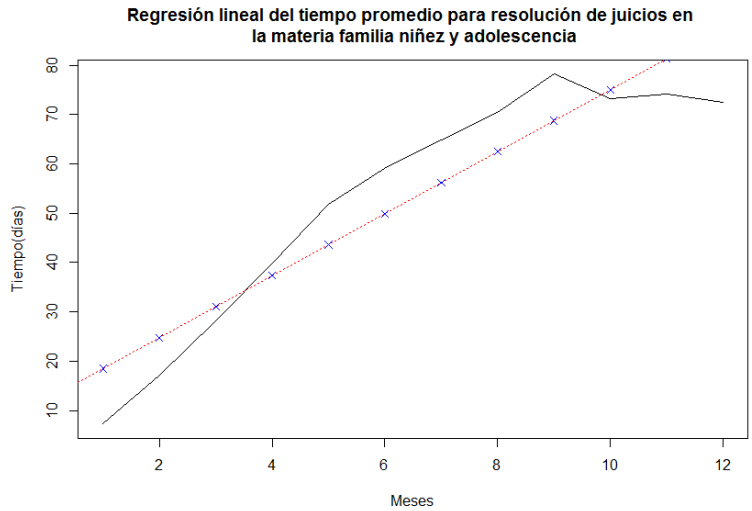

Figura 52. Regresión lineal del tiempo promedio para resolución de juicios en la materia familia niñez y adolescencia.

En la Figura 52, se observa la línea negra que representan los datos reales mientras que la roja representa los datos generados por la regresión lineal. Como antes ya se había mencionado, los juicios se resuelven entre un rango de 7 a 80 días. Pero si se visualiza la gráfica se puede notar (línea roja) que para el mes 12, se dispara totalmente. Esto implica que puede llevar a predecir un tiempo expresado en días totalmente fuera de la realidad. Además, se visualiza que para la regresión lineal, todo juicio tiene una constante entre 15 y 20 días mientras que en la realidad es inferior a los 10 días.

A continuación, la ecuación del modelo lineal:

$$Y = 12.21 + 6.3X$$

Figura 53. Ecuación lineal del tiempo promedio para resolución de juicios en la materia familia niñez y adolescencia.

Donde:

Y: número de días en que termina un juicio

X: los meses.

El modelo lineal entrega los siguientes resultados:

Tabla 8
Resumen del modelo lineal.

DATOS	
Intercepción	12.2074
Variable	6.2832
RMSE	8.4812
MAE	7.6583

En la Tabla 8, se observa valores que son las constantes y la pendiente. Y también, se puede notar el RMSE y MAE que no son muy diferentes entre ambos, pero existe una pequeña diferencia lo cual es aceptable para el modelo.

Regresión con el algoritmo SVM

Ahora se probará los datos con el algoritmo SVM para realizar la regresión y poder comparar con nuestra curva real. (Ver Anexo 7, Literal C. Script para generación de la gráfica).

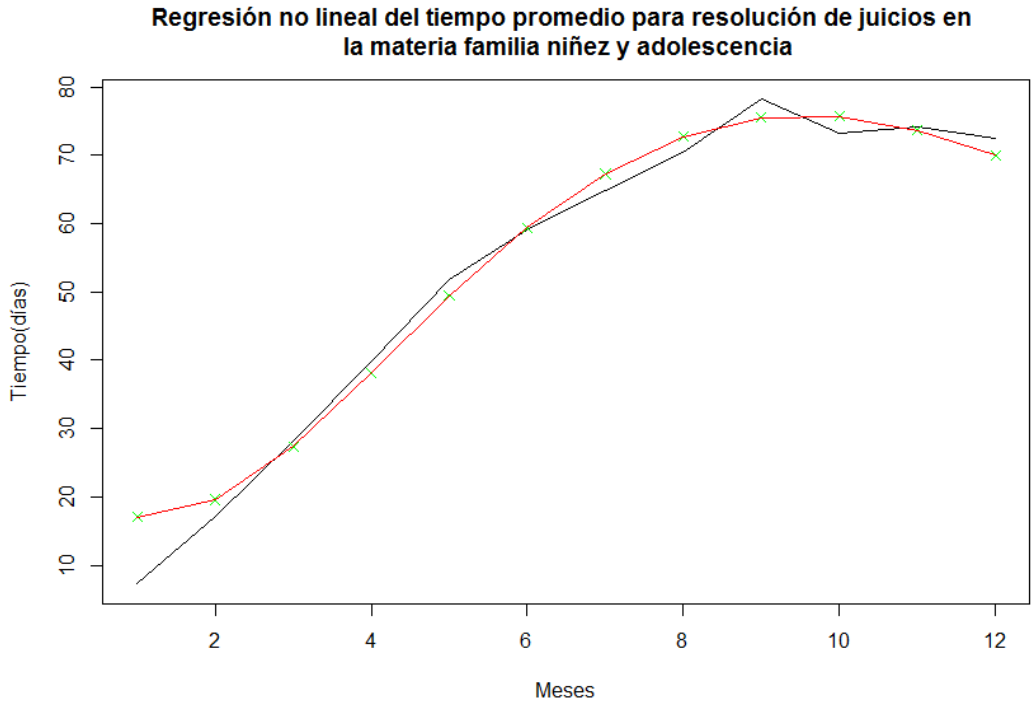

Figura 54. Regresión no lineal aplicando algoritmo SVM para el tiempo promedio para resolución de juicios en la materia familia niñez y adolescencia.

En la Figura 54 se tiene la gráfica de línea negra que son los datos reales mientras que la roja son los generados por el modelo no lineal de regresión aplicando el modelo SVM; se observa que las diferencias entre ambas curvas son casi idénticas. Para ser más técnicos se va a analizar los datos de la Tabla resumen para ver si es mejor o peor que el método lineal.

Tabla 9
Resumen del Modelo SVM.

DATOS	
Intercepción	NA
Variable	NA
RMSE	3.4383
MAE	2.5162

En la Tabla 9, se observa que el RMSE y el MAE son muy inferiores que el método lineal, casi 2.5 veces menores. Esto indica que el modelo es muy aproximado a lo real y por lo tanto dará resultados más aproximados.

- **Objetivo 3**. Determinar las tendencias de los procesos judiciales utilizando el historial de registros de los procesos judiciales del año 2015 en materia de familia niñez y adolescencia.

Para determinar la tendencia de los procesos judiciales en base a los datos que se tiene a disposición, se realizará utilizando la cantidad de juicios por unidad de mes. Así, de esta forma se tendrá la tendencia de resolución de juicios resueltos por mes y se podrá pronosticar el mes siguiente.

Regresión con el modelo LM

A continuación la siguiente gráfica de los casos resueltos respecto a los meses. (Ver Anexo 8, Literal B. Script para generación de la gráfica).

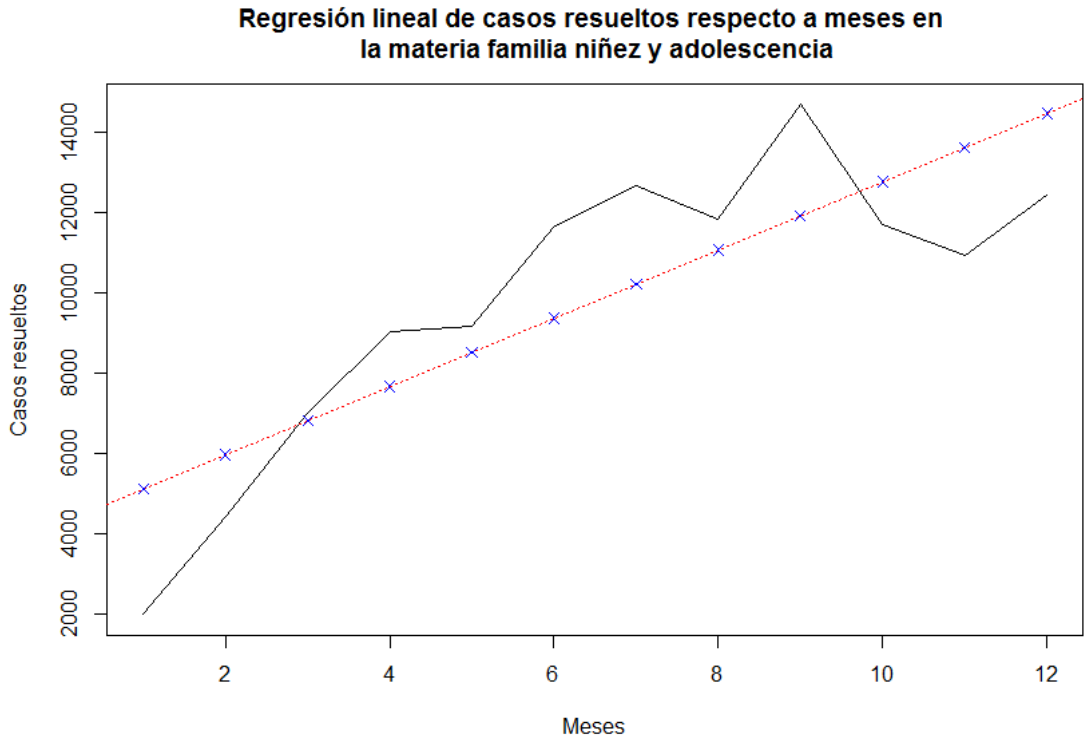

Figura 55. Regresión lineal de casos resueltos respecto a meses en la materia familia niñez y adolescencia.

En la Figura 55, se identifica dos curvas, la línea negra representa a los datos reales mientras que la roja son los puntos generados por el modelo lineal. Al observar ambas gráficas, se nota que como punto inicial inician en 2000 y la otra en 4500. Es una gran diferencia entre el modelo real y el generado por eso la sumatoria de los errores en RMSE y MAE son bastantes altos como lo describe la Tabla 10. También, se observa que el modelo lineal experimenta un alza al fin de año que está mucho más lejos de los datos reales.

A continuación, la ecuación del modelo lineal:

$$Y = 4270.2 + 848.5X$$

Figura 56. Ecuación lineal de casos resueltos respecto a meses en la materia familia niñez y adolescencia.

Donde:

Y: cantidad de juicios resueltos por mes

X: representa a los meses

El modelo lineal entregó los siguientes resultados:

Tabla 10
Resumen del modelo lineal.

DATOS	
Intercepción	4270.2
Variable	848.5
RMSE	1964.513
MAE	1742.453

Regresión con el algoritmo SVM

A continuación, se realizará una regresión no lineal aplicando el modelo SVM y así verificar los resultados para compararlos con el modelo de regresión lineal y seleccionar el mejor para la predicción. (Ver Anexo 8, Literal C. Script para generación de la gráfica).

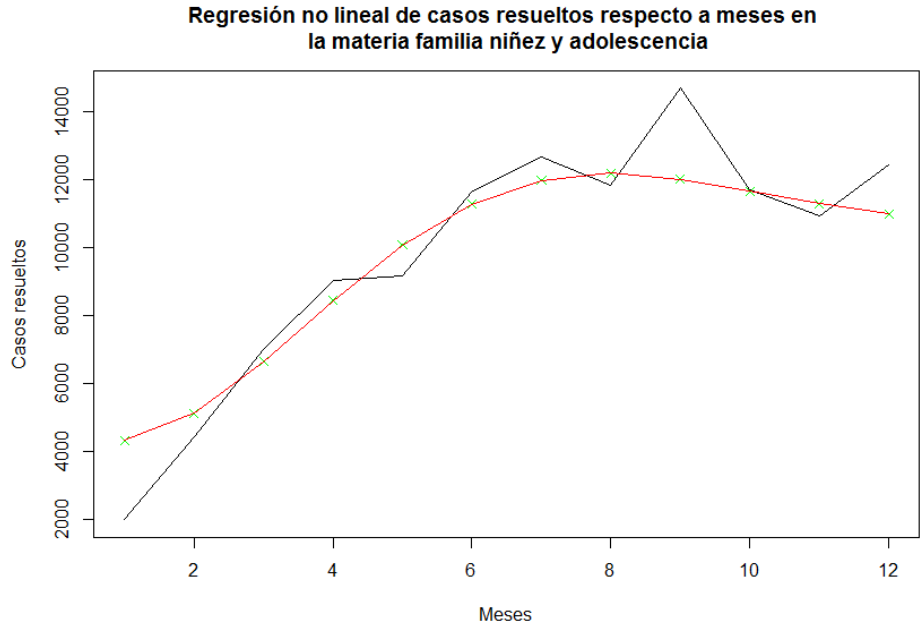

Figura 57. Regresión no lineal de casos resueltos respecto a meses en la materia familia niñez y adolescencia.

Si se observa la gráfica de la Figura 57, se notan dos curvas: la línea negra representa a los datos reales mientras que la roja representa al modelo no lineal SVM. Si se analiza la regresión del modelo generado, se nota que tiene el mismo inicio que el modelo lineal, es decir tienen un punto de inicio de 4000 pero diferente de allí en adelante en todos los puntos generados, porque el actual modelo se ajusta más a la curva real.

A continuación, la Tabla resume del modelo SVM.

Tabla 11
Resumen del modelo SVM.

DATOS	
Intercepción	NA
Variable	NA
RMSE	1197.804
MAE	899.5319

Al comparar la Tabla 11 del modelo SVM con la Tabla 10 del modelo LM la sumatoria de los errores de los residuos es totalmente inferior a la última. En otras palabras, el modelo SVM en todos los análisis realizados siempre tiene una cantidad mínima de errores hablando de forma general. Se dejará el análisis de los modelos para el capítulo de evaluación del modelado.

3.4.4 Evaluación del modelado

En esta actividad hay que tener claro que se orienta a analizar los objetivos definidos en la minería y no los del negocio. Se utilizará como punto de referencia las dos técnicas mencionados para medir el error de las predicciones de los modelos que se han probado.

Se comenzará explicando el primer modelo 1.1. del objetivo 1, se cuenta con un RMSE de 1994.134 para el modelo lineal mientras que el RMSE para el modelo SVM es de 1180.685, por lo tanto, el modelo SVM tiene una tasa de errores del 26% inferior al del modelo LM. También, se analizó el MAE, en el modelo LM tiene un valor de 1770.52 mientras que en el modelo SVM tiene un valor de 892.42 y la tasa de porcentaje de error entre ambos representa el 33%, lo que significa que el modelo LM tiene una tasa de porcentaje de error 33% más que el modelo SVM. Entonces, de acuerdo al análisis se selecciona al modelo SVM para resolver el objetivo 1.1.

Para el mismo objetivo 1, del modelo 1.2, se cuenta con un RMSE de 8.42 para el modelo lineal mientras que para el modelo SVM se tiene un RMSE de 3.46 lo cual es muy inferior al modelo lineal, en términos de porcentaje representa un 42% de errores más para el modelo lineal. Analizando el MAE en el modelo LM se tiene un valor de 7.59 mientras el MAE para el modelo SVM es de 2.53 lo que equivale a un 50% de errores menos que el método lineal. Por consecuencia, al encontrar que los mejores resultados favorecen al método SVM, se seleccionara a este modelo para resolver el objetivo 1 del modelo 1.2.

Para el segundo modelo del objetivo 2, se cuenta con RMSE del modelo lineal de 8.48 mientras para el SVM es de 3.43 lo que significa una tasa de error del 42% menos para el modelo SVM. También, se comparó el MAE del lineal con un valor de 7.66 y para el SVM con un valor de 2.53 que representa un porcentaje de error del 50% menos en el modelo SVM. Entonces, de acuerdo a los análisis, se verifico que el mejor modelo para resolver el objetivo 2 es el modelo SVM.

Por último, se tiene el tercer modelo para el objetivo 3, en el que se ha obtenido para el modelo lineal un RMSE de 1964.51 mientras para el modelo SVM un valor de 1197.8 lo que representa una tasa del 24% de error adicional en el modelo lineal. Además, se analizó el MAE del modelo lineal con un valor de 1742.45 y para el SVM con un valor de 899.53 lo que representa una tasa de error del 32% más para el modelo lineal. En base a estas evidencias se empleará el modelo SVM para resolver el objetivo 3.

En la siguiente Tabla se puede observar los valores para los distintos indicadores para hacer una mejor comparativa:

Tabla 12
Resumen de errores de los modelos.

	Error absoluto Medio(MAE)		**Error Cuadrático Medio(RMSE)**	
	SVM	LM	SVM	LM
Modelo 1.1	892.42	1770.52	1180.69	1994.13
Modelo 1.2	2.53	7.58	3.46	8.42
Modelo 2	2.52	7.66	3.44	8.48
Modelo 3	899.53	1742.45	1197.8	1964.51

3.5 Evaluación

En CRISP-DM sugiere que se debe decir si es factible o no el uso de un modelo para solventar el objetivo del negocio. En caso de estar mal definido o no es posible lograr el objetivo se tendrá que aclarar que no hay modelo que pueda ser empleado para dicho objetivo.

3.5.1 Evaluación de los resultados

En esta actividad se decidirá si el modelo cubre la necesidad del objetivo en caso de no cumplir se lo omitirá.

- **Objetivo 1**

Modelo 1.1: De acuerdo al punto 4.4.4 se observa que el análisis del modelo del SVM está muy ajustado a los datos reales y los patrones descubiertos permitirán colocar mayor personal para atender en dichas judicaturas y al mismo tiempo se podrán hacer predicciones de los casos resueltos por mes en los tipos de acciones que guardan relaciones como: contencioso general, deprecatorio, especial, ordinario y

verbal sumario. Por lo tanto, se lo considera aceptable desde el punto de vista de los objetivos del negocio.

Modelo 1.2: Al igual que el modelo 1.1 este cumple con las expectativas del objetivo porque lo que se ha descubierto permitirá hacer predicciones del tiempo en que demora un juicio de acuerdo al mes en los tipos de acciones relacionados como: contencioso general, deprecatorio, especial, ordinario y verbal sumario. Por lo tanto, se considera que el modelo descubierto para este objetivo es aceptable.

- **Objetivo 2**

Los resultados de las observaciones (4.4.4) y predicciones simuladas arrojaron resultados bastantes ajustados a la realidad por tal motivo se considera que el modelo es aceptable porque permitirá realizar predicciones del tiempo promedio en que demora la resolución de un juicio en la materia de la familia niñez y adolescencia permitiendo saber por anticipado a los ciudadanos el tiempo estimado y a la función judicial si debe o no maximizar sus esfuerzos para dar un mejor servicio ya que a eso se quiere llegar.

- **Objetivo 3**

Predecir la cantidad de procesos judiciales resueltos dado un mes específico y que el modelo se ajuste a los datos reales, se considera aceptable tanto desde el punto técnico y práctico para la institución.

Modelos aprobados

Dado los argumentos en cada uno de los objetivos tanto desde la minería y del negocio se ha procedido a aprobar todos los modelos de cada uno de los objetivos por su extrema importancia para la institución.

3.5.2 Revisión

Como todos los objetivos se han podido demostrar con éxito no hay necesidad de realizar revisión alguna.

3.5.3 Determinar los próximos pasos

Lo que se realizará a continuación es el proceso de implantación de los modelos obtenidos en cada uno de los objetivos. Pero solo se describirá lo que se debe realizar porque aún no se autorizado implementar los modelos en la institución.

3.6 Implantación

Para tener una idea general de cuando se implemente los modelos obtenidos, en esta sección se debe definir o aclarar cómo el programa tendrá acceso a los datos de la base de datos, a que aplicativos se deberá transmitir los resultado y cuánto tiempo se deberá entrenar de nuevo el modelo para tener ajustes a los modelo.

3.6.1 Planear la implantación

Para llevar a cabo la implementación de los modelos es necesario que se guarde el modelo entrenado en la base de datos y se establezca las versiones de cada uno de ellos. Además, se debe tener instalado el programa R, se recomienda que sea un servidor a parte porque el cálculo puede que afecte la latencia de la base y haga que se caiga el sistema.

Otro punto a considerar, es la base de datos, afortunadamente R tiene un driver para conectarse a la base que se utiliza en la institución y no dará problemas. Sin embargo, el mayor problema a resolver es como trasmitir los resultados a cada uno de los usuarios que la soliciten, para esto se debe instalar un paquete que hace que R actué como un servidor web y de esta forma a través de un servicio rest se llevará la petición al cliente.

3.6.2 Planear la Monitorización y Mantenimiento

Monitorizar y dar mantenimiento a modelos que se encuentran constantemente ajustándose en base a entradas que cambian con frecuencia no es nada sencillo así que se definirá los pasos que se deben seguir.

- Trabajar sobre una base de datos especializada solo para el análisis de minería de datos.
- Replicar la base transaccional a la base de minería en las noches.
- Detener R cuando se estén realizando las réplicas de la base de datos.
- Realizar copias mensuales de la base donde se está realizando la minería.
- Guardar los resultados de las predicciones para luego aplicar técnicas estadistas para tratar de encontrar datos atípicos y corregir el modelo si sucediera.

3.7 Producir el informe final

Durante la definición de los objetivos de la minería de datos se observó que tanto el departamento de estadística y gestión procesal tenían plena consciencia por descubrir las relaciones o patrones que estos datos ocultaban, fue así entonces cuando se empezó con la exploración de los datos donde se descubrió que había juicios que su resolución tomo ni un día, lo cual no tenía lógica.

Posteriormente, se realizó un histograma a cada uno de los tipos de acción y fue allí donde se observó que los tipos de acciones como: contencioso general, deprecatorio, especial, ordinario y verbal sumario; con mayor demanda.

Luego, los diagramas de cajas revelaron que existían datos atípicos dentro del conjunto de datos por lo que se debía tomar medidas porque estos alterarían los resultados, en este caso se decidió ignorarlos porque solo representaban el 2% de los registros.

Por otra parte, en el desarrollo de los objetivos, se tiene que el más laborioso de todos fue el primero porque se debía encontrar asociaciones o relaciones que sean de interés para la institución dentro de los datos. Pero se estaba pensando que en caso de

no encontrar ningún patrón simplemente se da por alto ese objetivo. Afortunadamente, se identificó los tipos de acciones que mantenían estas características que se buscaba y se pudo definir los modelos.

Otro de los puntos más laboriosos fue el escribir el código y entender los modelos que vienen en los paquetes que R proporciona. Al inicio se tuvo que realizar varias pruebas y errores hasta obtener los resultados que se presentan en esta investigación, pero conforme se avanzaba se encontró que la herramienta proporcionaba tanta flexibilidad para predecir modelos que otras herramientas sería difícil realizar.

Sin embargo, el proceso más tedioso, pero de gran importancia, fue la parte de evaluación de los modelos tanto orientado a los objetivos de la minería y del negocio. Aquí se evaluó los errores de MSE y RMSE obtenidos en cada objetivo para de esta forma seleccionar el modelo a ser implementado en la institución. Durante el proceso de evaluación y selección del modelo se observó que los modelos SVM mostraron excelentes resultados a la hora de predecir, por lo tanto, fue el modelo ganador.

Finalmente, después de la evaluación de los modelos se llegó a la conclusión que se satisfacían los tres objetivos del negocio y también se describió las tareas que se deben realizar para poder implementar este modelo en la institución (Actividad de implantación).Se debe aclarar que el informe es de forma general, el que se presentará para la presentación contendrá imágenes de los resultados y serpa detallado minuciosamente.

3.8 Revisar el proyecto

Como toda actividad final de una fase de CRISP-DM presenta una línea de retorno hacia otras fases posteriores, en caso de que en la investigación se tenga que mejorar algo o revisar algún proceso, se tendrá que ir y solucionarlo.

Se tiene que admitir que la parte más laboriosa es el de realizar mediante script e investigar el uso de las funciones de modelado en R por lo que se extendió un poco el tiempo de la culminación de la exploración de los datos. Sin embargo, se tuvo puntos

positivos porque la herramienta al ser muy flexible permitiendo realizar multigráficas de modelos cosas que en otros programas tomaría mucho tiempo en realizarlos. A parte de esto, se realizaron funciones que permitieron ahorrar código y una vez que se sabe cómo implementar las funciones de modelado, es muy sencillo aplicarlos.

CAPÍTULO IV
CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- La metodología CRISP-DM es una guía muy detallada que permite llevar el control y alcanzar de los objetivos de la minería de datos.
- El programa R es una herramienta estadística muy poderosa para la exploración y análisis de los datos.
- Durante la exploración de datos se encontró una forma muy óptima de encontrar los datos atípicos que es mediante los diagramas de cajas.
- La corrección de los datos atípicos permitió tener predicciones más acertadas a la realidad.
- La limpieza de datos y selección de los mismos son de mucha importancia para las predicciones.
- Los histogramas empleados para la exploración de los datos fueron de mucha utilidad para determinar el tiempo promedio de proceso judicial en la materia familia niñez y adolescencia.
- Se descubrió que los tipos de acciones con mayor demanda en la materia familia niñez y adolescencia son: contencioso general, deprecatorio, especial, ordinario y verbal sumario.
- Así mismo, también se observó que los meses que experimentan menor demanda son de enero hasta mayo luego la demanda aumenta hasta que se mantiene hasta diciembre.
- Se pudo determinar la tendencia de los procesos judiciales a través de los modelos SVM y LM en la materia familia niñez y adolescencia.
- Se pudo determinar el tiempo promedio para la finalización de un proceso judicial en la materia familia niñez y adolescencia.
- Durante la investigación se llegó a la conclusión que encontrar un modelo matemático para representar el comportamiento de la tendencia de los datos estaba muy alejado de la realidad.

- Se llegó a la conclusión que el modelo SVM es el que mayor se ajustó al análisis permitiendo ser el seleccionado para las predicciones.

4.2 Recomendaciones

- Se aconseja que la persona que manipule los modelos sea el que desarrollo los mismos, de caso contrario debería ser capacitado, es con el objetivo de evitar una mala interpretación del modelo.
- Se recomienda utilizar una conexión directa a una copia de la base de datos o data Warehouse debido al constante cambio de los datos o manipulación de la información.
- Se recomienda ajustar los modelos en caso de ser necesario.
- Se recomienda usar los modelos SVM para las predicciones porque son los más precisos.
- Sería aconsejable obtener un respaldo semestral de los datos para su análisis por motivos de seguridad.
- Si se desea implementar este servicio para aplicativos móviles es necesario de crear un servicio rest para el consumo del modelo.
- Para examinar grandes cantidades de información en el orden de terabytes es necesario tener equipos con una gran cantidad de RAM y un enorme poder de procesamiento.

BIBLIOGRAFÍA

Barrientos, F., & Ríos, S. (2013). Aplicación de minería de datos para predecir fuga de clientes en la industria de las telecomunicaciones. *Revista de Ingenieria de Sistemas*, 75-77.

Barrios, M. (2010). Modelo del Negocio. *Americana*.

Brito, P. (2010). *Objetivos de Negocio y Procesos de Minería de Datos Basados en Sistemas Inteligentes.* Argentina.

C, S. (2000). *el modelo CRISP-DM: el nuevo plan para la minería de datos.* The Journal of Data Warehousing.

CEI. (2010). *Manual Básico para Elaborar Plan de Negocio para PYMEs.* Nicaragua.

F., T. M. (2004). *Estadística.* Mexico: Pearson Educación.

Fayyad, U., & Haussler, D. S. (1996). Mining Scientific Data. *Communications of the ACM*, 51-57.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge and data mining.* Cambridge (Massachussets): AAAI/MIT Press. .

Folgueiras Bertomeu, P. (2010). *Métodos y técnicas de recogida y análisis de.* Argentina: Universidad de Barcelona.

Función Judicial. (12 de 03 de 2016). Obtenido de http://www.funcionjudicial.gob.ec/index.php

Función Judicial(Antecedentes). (13 de 03 de 2016). Obtenido de http://www.funcionjudicial.gob.ec/www/pdf/informatica/PLANESTRATEGICOOPERATIVODNI-CJ.pdf

Galán Cortina, V. (2015). *APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO.* Madrid: Escuela Politécnica Superior Ingeniería en Informática .

Gommez Díaz, H., & Cerón Reyes, M. d. (2010). *Minería de datos.*

Grossman, R. L., Hornik, M., & G., M. (2012). Data mining standars initiatives. *Communications of ACM*, 59-61.

IBM. (2012). *Manual CRISP-DM deIBM SPSS Modeler.* Estados Unidos.

José Hernández Orallo, M. J. (2004). *Introducción a la Minería de Datos. Ed. Pearson Educación.* Pearson Educación.

KDnuggets. (15 de 12 de 2016). Obtenido de www.kdnuggets.com: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

K-Means, s.f. (16 de 04 de 2016). Obtenido de wikipedia: https://es.wikipedia.org/wiki/K-means

Llavona Arregui, J. L. (2010). *Terminología de Estadística y Minería de Datos en Lengua Inglesa.* Madrid, España: ISBN.

Moine, J. M., Haedo, A. S., & & Gordillo, S. (2012). *Estudio Corporativo de los Datos.* Argentina: UTN Rosario.

Molero Castillo, G. G. (2008). *Desarrollo de un Modelo basado en técnicas de Minería de Datos para clasificar zona climatológicamente similares en el estado de Michoacán.* México: Universidad Nacional Autónoma de México.

Montserrat, P. V. (2001). Series Temporales. En P. V. Montserrat, *Series Temporales* (págs. 67-68). Barcelona, España: Edición de la Universidad Politécnica de Cataluña.

Moral Peláez, I. (2012). *Modelos de regresión:Llineal Simple y Regresión Logogística.* España.

Peralta Cochancela, D. E. (2009). *Proyecto de Minería de Datos para el Annálisis del Comportamiento de los Clientes de Telecomunicaciones.* Cuenca: Universidad Politecnica Salesiana Sede Cuenca.

Pete Chapman, J. C. (2000). *CRISP-DM 1.0, Step-by-step Data Mining Guide.*

Red Neuronal, s. f. (16 de 05 de 2016). Obtenido de wikipedia: https://es.wikipedia.org/wiki/Red_neuronal

Reglas de Asociación, s.f. (12 de 04 de 2016). Obtenido de wikipedia: https://es.wikipedia.org/wiki/Reglas_de_asociaci%C3%B3n

Regresión Lineal, s.f. (20 de 05 de 2016). Obtenido de wikipedia: https://es.wikipedia.org/wiki/An%C3%A1lisis_de_la_regresi%C3%B3n

Revista Judicial de Derecho. (13 de 03 de 2016). Obtenido de http://www.derechoecuador.com/articulos/detalle/archive/doctrinas/funcionjudicial/2005/11/24/que-es-la-funcion-judicial

Rigeiro. (2012). Procesamiento y análisis de los Datos.

Rodas, J. (2001). Un Ejercicio de análisis utilizando rough sets en un dominio de educación superior mediante el proceso KDD. *Barcelona: Departamento de Lenjuages y Sistemas Informáticos, Universidad Politécnica de Cataluña*.

Rojas, D. O. (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* Costa Rica.

Rokach, L., & Maimon, O. (2008). Data mining with decision trees: theory and applications. *World Scientific*.

Romeau Guallart, P. (2010). *Minería de Datos Aplicada Al Análisis del Tratamiento Informativo de la Drogadicción.* Moncada: CEU.

Anexo 1: Glosario de Terminología de Minería de Datos

- Análisis de series de tiempo (time-series): Análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominante de los datos.
- Análisis prospectivo de datos: Análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históricos.
- Análisis exploratorio de datos: Uso de técnicas estadísticas tanto gráficas como descriptivas para aprender acerca de la estructura de un conjunto de datos.
- Clustering (agrupamiento): Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles.
- Data cleansing: Proceso de asegurar que todos los valores en un conjunto de datos sean consistentes y correctamente registrados.
- Data Mining: La extracción de información predecible escondida en grandes bases de datos.
- Datos anormales: Datos que resultan de errores o que representan eventos inusuales.
- Modelo analítico: Una estructura y proceso para analizar un conjunto de datos. Por ejemplo, un árbol de decisión es un modelo para la clasificación de un conjunto de datos.
- Modelo lineal: Un modelo analítico que asume relaciones lineales entre una variable seleccionada (dependiente) y sus predictores (variables independientes).
- Modelo no lineal: Un modelo analítico que no asume una relación lineal en los coeficientes de las variables que son estudiadas.
- Modelo predictivo: Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos.
- Outlier: Un ítem de datos cuyo valor cae fuera de los límites que encierran a la mayoría del resto de los valores correspondientes de la muestra. Puede indicar

datos anormales. Deberían ser examinados detenidamente, pueden dar importante información.

- Regresión lineal: Técnica estadística utilizada para encontrar la mejor relación lineal que encaja entre una variable seleccionada (dependiente) y sus predicados (variables independientes).
- RMSE: Error cuadrático medio y sirve para evaluar un modelo de regresión
- MAE: Error absoluto medio y sirve para evaluar la calidad en los modelos de regresión.

Anexo 2: Exploración de datos

a. Script en R para creación de histogramas de todos los meses del 2015.

```
df <- read.table("analisis_2.csv",header = TRUE,sep = ";")
enero<-subset(df,as.Date(df$Fecha) < '2015-02-01' )
febrero<-subset(df,as.Date(df$Fecha) >= '2015-02-01' & as.Date(df$Fecha) <= '2015-02-28')
marzo<-subset(df,as.Date(df$Fecha) >= '2015-03-01' & as.Date(df$Fecha) <= '2015-03-31' )
abril<-subset(df,as.Date(df$Fecha) >= '2015-04-01' & as.Date(df$Fecha) <= '2015-04-30' )
mayo<-subset(df,as.Date(df$Fecha) >= '2015-05-01' & as.Date(df$Fecha) <= '2015-05-31' )
junio<-subset(df,as.Date(df$Fecha) >= '2015-06-01' & as.Date(df$Fecha) <= '2015-06-30' )

julio<-subset(df,as.Date(df$Fecha) >= '2015-07-01' & as.Date(df$Fecha) <= '2015-07-31' )
agosto<-subset(df,as.Date(df$Fecha) >= '2015-08-01' & as.Date(df$Fecha) <= '2015-08-31' )
sept<-subset(df,as.Date(df$Fecha) >= '2015-09-01' & as.Date(df$Fecha) <= '2015-09-30' )
oct<-subset(df,as.Date(df$Fecha) >= '2015-10-01' & as.Date(df$Fecha) <= '2015-10-31' )
nov<-subset(df,as.Date(df$Fecha) >= '2015-11-01' & as.Date(df$Fecha) <= '2015-11-30' )
dic<-subset(df,as.Date(df$Fecha) >= '2015-12-01' & as.Date(df$Fecha) <= '2015-12-31' )

par(mfrow=c(4,3))
hist(enero$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos resueltos",main="Enero")
```

```
hist(febrero$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Febrero")
hist(marzo$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Marzo")
hist(abril$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Abril")
hist(mayo$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Mayo")
hist(junio$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Junio")

hist(julio$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Julio")
hist(agosto$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Agosto")
hist(sept$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Septiembre")
hist(oct$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Octubre")
hist(nov$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Noviembre")
hist(dic$Tiempo,breaks=60, col="red",xlab="N° días",ylab="Casos
resueltos",main="Diciembre")
```

b. Script en R para creación de tendencias de juicios resueltos por mes del 2015.

```
library(ggplot2)
library(scales)
library(grid)

df <- read.table("analisis_2.csv",header = TRUE,sep = ";")
enero<-subset(df,as.Date(df$Fecha) < '2015-02-01' )
febrero<-subset(df,as.Date(df$Fecha) >= '2015-02-01' & as.Date(df$Fecha) <=
'2015-02-28')
marzo<-subset(df,as.Date(df$Fecha) >= '2015-03-01' & as.Date(df$Fecha) <= '2015-
03-31' )
abril<-subset(df,as.Date(df$Fecha) >= '2015-04-01' & as.Date(df$Fecha) <= '2015-
04-30' )
mayo<-subset(df,as.Date(df$Fecha) >= '2015-05-01' & as.Date(df$Fecha) <= '2015-
05-31' )
junio<-subset(df,as.Date(df$Fecha) >= '2015-06-01' & as.Date(df$Fecha) <= '2015-
06-30' )

julio<-subset(df,as.Date(df$Fecha) >= '2015-07-01' & as.Date(df$Fecha) <= '2015-
```

```
07-31' )
agosto<-subset(df,as.Date(df$Fecha) >= '2015-08-01' & as.Date(df$Fecha) <= '2015-
08-31' )
sept<-subset(df,as.Date(df$Fecha) >= '2015-09-01' & as.Date(df$Fecha) <= '2015-
09-30' )
oct<-subset(df,as.Date(df$Fecha) >= '2015-10-01' & as.Date(df$Fecha) <= '2015-10-
31' )
nov<-subset(df,as.Date(df$Fecha) >= '2015-11-01' & as.Date(df$Fecha) <= '2015-
11-30' )
dic<-subset(df,as.Date(df$Fecha) >= '2015-12-01' & as.Date(df$Fecha) <= '2015-12-
31' )


par(mfrow=c(2,2))
a1 <- ggplot(enero, aes(as.Date(enero$Fecha),enero$Tiempo,group="60")) +
geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Enero") + ylab("Casos
resueltos")+theme_bw()

a2<- ggplot(febrero, aes(as.Date(febrero$Fecha),febrero$Tiempo,group="60")) +
geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Feb.") + ylab("Casos
resueltos")+theme_bw()

a3<- ggplot(marzo, aes(as.Date(marzo$Fecha),marzo$Tiempo,group="60")) +
geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Marzo") + ylab("Casos
resueltos")+theme_bw()

a4<- ggplot(abril, aes(as.Date(abril$Fecha),abril$Tiempo,group="60")) +
geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Abril") + ylab("Casos
resueltos")+theme_bw()

a5<- ggplot(mayo, aes(as.Date(mayo$Fecha),mayo$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Mayo") + ylab("Casos
resueltos")+theme_bw()

a6<- ggplot(junio, aes(as.Date(junio$Fecha),junio$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Junio") + ylab("Casos
resueltos")+theme_bw()

a7<- ggplot(julio, aes(as.Date(julio$Fecha),julio$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Julio") + ylab("Casos
resueltos")+theme_bw()

a8<- ggplot(agosto, aes(as.Date(agosto$Fecha),agosto$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Agosto") + ylab("Casos
resueltos")+theme_bw()
```

```
a9<- ggplot(sept, aes(as.Date(sept$Fecha),sept$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Sept.") + ylab("Casos
resueltos")+theme_bw()

a10<- ggplot(oct, aes(as.Date(oct$Fecha),oct$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Oct.") + ylab("Casos
resueltos")+theme_bw()

a11<- ggplot(nov, aes(as.Date(nov$Fecha),nov$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Nov.") + ylab("Casos
resueltos")+theme_bw()

a12<- ggplot(dic, aes(as.Date(dic$Fecha),dic$Tiempo)) + geom_line() +
  scale_x_date(labels = date_format("%m-%d"))+ xlab("Dic.") + ylab("Casos
resueltos")+theme_bw()


vplayout <- function(x, y) viewport(layout.pos.row = x, layout.pos.col = y)

grid.newpage()
pushViewport(viewport(layout = grid.layout(4, 3)))

print(a1, vp = vplayout(1,1))
print(a2, vp = vplayout(1,2))
print(a3, vp = vplayout(1,3))

print(a4, vp = vplayout(2,1))
print(a5, vp = vplayout(2,2))
print(a6, vp = vplayout(2,3))

print(a7, vp = vplayout(3,1))
print(a8, vp = vplayout(3,2))
print(a9, vp = vplayout(3,3))

print(a10, vp = vplayout(4,1))
print(a11, vp = vplayout(4,2))
print(a12, vp = vplayout(4,3))
```

c. Script en R para generación de histograma del 2015

```
df <- read.table("analisis_2.csv",header = TRUE,sep = ";")

hist(df$Tiempo,xlim=c(1,400),density=20,breaks = 20,main="Histograma de
Procesos Judiciales resueltos en materia \n Familia Niñez y Adolescencia", xlab =
"N° días", ylab="Juicios resueltos")
```

d. Script en R para generación de tendencias de resoluciones judiciales

```
df <- read.table("analisis_2.csv",header = TRUE,sep = ";")

plot(df$Tiempo,main="Juicios resueltos vs N° días", xlab = "Juicios resueltos ",
ylab="N° días")
```

e. Sript en R para generación de curva de histograma

```
df <- read.table("analisis_2.csv",header = TRUE,sep = ";")
 h<-hist(df$Tiempo, breaks=60, col="red", xlab="N° de días",ylab="Frecuencia",
        main="Distribución normal de los juicios resueltos en materia \n Familia Niñez
y Adolescencia")

 xfit<-seq(min(df$Tiempo),max(df$Tiempo),length=60)
 yfit<-dnorm(xfit,mean=mean(df$Tiempo),sd=sd(df$Tiempo))
 yfit <- yfit*diff(h$mids[1:2])*length(df$Tiempo)
 lines(xfit, yfit, col="blue", lwd=2)
```

Anexo 3: Outlier

a. Script en R para detección de datos atípicos

```
library(outliers)
library(lattice)
df <- read.table("analisis_2.csv",header = TRUE,sep = ";")
enero<-subset(df,as.Date(df$Fecha) < '2015-02-01'  ) # & df$Tiempo <=24
febrero<-subset(df,as.Date(df$Fecha) >= '2015-02-01' & as.Date(df$Fecha) <=
'2015-02-28')
marzo<-subset(df,as.Date(df$Fecha) >= '2015-03-01' & as.Date(df$Fecha) <= '2015-
03-31' )
abril<-subset(df,as.Date(df$Fecha) >= '2015-04-01' & as.Date(df$Fecha) <= '2015-
04-30' )
mayo<-subset(df,as.Date(df$Fecha) >= '2015-05-01' & as.Date(df$Fecha) <= '2015-
05-31' )
junio<-subset(df,as.Date(df$Fecha) >= '2015-06-01' & as.Date(df$Fecha) <= '2015-
06-30' )
julio<-subset(df,as.Date(df$Fecha) >= '2015-07-01' & as.Date(df$Fecha) <= '2015-
07-31' )
agosto<-subset(df,as.Date(df$Fecha) >= '2015-08-01' & as.Date(df$Fecha) <= '2015-
08-31' )
sept<-subset(df,as.Date(df$Fecha) >= '2015-09-01' & as.Date(df$Fecha) <= '2015-
09-30' )
oct<-subset(df,as.Date(df$Fecha) >= '2015-10-01' & as.Date(df$Fecha) <= '2015-10-
```

```
31' & df$Tiempo < 280)
nov<-subset(df,as.Date(df$Fecha) >= '2015-11-01' & as.Date(df$Fecha) <= '2015-
11-30' & df$Tiempo < 260)
dic<-subset(df,as.Date(df$Fecha) >= '2015-12-01' & as.Date(df$Fecha) <= '2015-12-
31' & df$Tiempo < 250)
 par(mfrow=c(2,3))
 # boxplot(enero$Tiempo, main="Enero")
 # boxplot(febrero$Tiempo, main="Febrero")
 # boxplot(marzo$Tiempo, main="Marzo")
 # boxplot(abril$Tiempo, main="Abril")
 # boxplot(mayo$Tiempo, main="Mayo")
 # boxplot(junio$Tiempo, main="Junio")

 boxplot(julio$Tiempo, main="Julio")
 boxplot(agosto$Tiempo, main="Agosto")
 boxplot(sept$Tiempo, main="Septiembre")
 boxplot(oct$Tiempo, main="Octubre")
 boxplot(nov$Tiempo, main="Noviembre")
 boxplot(dic$Tiempo, main="Diciembre")
```

b. Script en R para visualización de corrección de datos atípicos

```
library(outliers)
library(lattice)
df <- read.table("analisis_2.csv",header = TRUE,sep = ";")

enero<-subset(df,as.Date(df$Fecha) < '2015-02-01'  ) # & df$Tiempo <=24
febrero<-subset(df,as.Date(df$Fecha) >= '2015-02-01' & as.Date(df$Fecha) <=
'2015-02-28')
marzo<-subset(df,as.Date(df$Fecha) >= '2015-03-01' & as.Date(df$Fecha) <= '2015-
03-31' )
abril<-subset(df,as.Date(df$Fecha) >= '2015-04-01' & as.Date(df$Fecha) <= '2015-
04-30' )
mayo<-subset(df,as.Date(df$Fecha) >= '2015-05-01' & as.Date(df$Fecha) <= '2015-
05-31' )
junio<-subset(df,as.Date(df$Fecha) >= '2015-06-01' & as.Date(df$Fecha) <= '2015-
06-30' )
julio<-subset(df,as.Date(df$Fecha) >= '2015-07-01' & as.Date(df$Fecha) <= '2015-
07-31' )
agosto<-subset(df,as.Date(df$Fecha) >= '2015-08-01' & as.Date(df$Fecha) <= '2015-
08-31' )
sept<-subset(df,as.Date(df$Fecha) >= '2015-09-01' & as.Date(df$Fecha) <= '2015-
09-30' )
oct<-subset(df,as.Date(df$Fecha) >= '2015-10-01' & as.Date(df$Fecha) <= '2015-10-
31' & df$Tiempo < 280)
nov<-subset(df,as.Date(df$Fecha) >= '2015-11-01' & as.Date(df$Fecha) <= '2015-
11-30' & df$Tiempo < 260)
```

```
dic<-subset(df,as.Date(df$Fecha) >= '2015-12-01' & as.Date(df$Fecha) <= '2015-12-31' & df$Tiempo < 250)

par(mfrow=c(2,3))
boxplot(agosto$Tiempo, main="Agosto")
boxplot(sept$Tiempo, main="Septiembre")
boxplot(oct$Tiempo, main="Octubre")
boxplot(nov$Tiempo, main="Noviembre")
boxplot(dic$Tiempo, main="Diciembre")
```

Anexo 4: Script de limpieza de datos

```
insert into
analisis_judicial("IdMateria","IdTipoAccion","NumDias","FechaResolucion")
select "IdMateria" ,"IdTipoAccion","NumDias", "FechaResolucion" from
proceso_judicial  where "IdMateria"='6'  and "NumDias" > 0 and  "NumDias" <=250
and "FechaResolucion"<= '2015-12-30'
```

Anexo 5: Tabla de tipos de acción de la materia familia niñez y adolescencia

Familia Niñez y Adolescencia	
6001	ADOLESCENTE INFRACTOR
6002	COMISION
6003	CONTENCIOSO GENERAL
6004	DEPRECATORIO
6005	DILIGENCIA PREPARATORIA
6006	ESPECIAL
6007	EXHORTO
6008	ORDINARIO
6009	SUMARIO
6010	VERBAL SUMARIO
6011	RECUSACIÓN
6012	SOLICITUDES
6014	MEDIDAS DE PROTECCION

Anexo 6: Scripts del objetivo 1

a. Script en R por tipo de acción en la familia niñez y adolescencia

```
library(ggplot2)
library(scales)
library(grid)

df <- read.table("analisis_obj1.csv",header = TRUE,sep = ";")

DATA6001<-subset(df,df$IdTA == 6001 & df$Mes <= 13 ) #adolescente infractor
```

```
DATA6002<-subset(df,df$IdTA == 6002 & df$Mes <= 6 ) #comision vacio
DATA6003<-subset(df,df$IdTA == 6003 & df$Mes <= 13 ) #contencioso general
DATA6004<-subset(df,df$IdTA == 6004 & df$Mes <= 13 ) #DEPRECATORIO
DATA6005<-subset(df,df$IdTA == 6005 & df$Mes <= 6 ) #DILIGENCIA PREPARATORIA
DATA6006<-subset(df,df$IdTA == 6006 & df$Mes <= 13 ) #especial
DATA6007<-subset(df,df$IdTA == 6007 & df$Mes <= 6 ) #EXHORTO

DATA6008<-subset(df,df$IdTA == 6008 & df$Mes <= 13 ) #ORDINARIO
DATA6009<-subset(df,df$IdTA == 6009 & df$Mes <= 6 ) #sumario
DATA6010<-subset(df,df$IdTA == 6010 & df$Mes <= 13 ) # verbal sumario
DATA6011<-subset(df,df$IdTA == 6011 & df$Mes <= 6 ) # recusación
DATA6012<-subset(df,df$IdTA == 6012 & df$Mes <= 6 ) # solicitudes
DATA6014<-subset(df,df$IdTA == 6014 & df$Mes <= 6 ) # medidas de protección



a1 <- ggplot(DATA6001,
aes(as.Date(DATA6001$Fecha),DATA6001$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ ggtitle("Adolescente infractor")
+xlab("") + ylab("")

a2 <- ggplot(DATA6002,
aes(as.Date(DATA6002$Fecha),DATA6002$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ ggtitle("Comisión") +xlab("") +
ylab("")


a3<- ggplot(DATA6003,
aes(as.Date(DATA6003$Fecha),DATA6003$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ ggtitle("Contencioso general")
+xlab("") + ylab("")

a4<- ggplot(DATA6004,
aes(as.Date(DATA6004$Fecha),DATA6004$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ ggtitle("Deprecatorio") +xlab("") +
ylab("")

a5<- ggplot(DATA6005,
aes(as.Date(DATA6005$Fecha),DATA6005$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ ggtitle("Diligencia preparatoria")
+xlab("") + ylab("")

a6<- ggplot(DATA6006,
```

```
aes(as.Date(DATA6006$Fecha),DATA6006$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Especial") +xlab("") +
ylab("")

a7<- ggplot(DATA6007,
aes(as.Date(DATA6007$Fecha),DATA6007$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Exhorto") +xlab("") +
ylab("")

a8<- ggplot(DATA6008,
aes(as.Date(DATA6008$Fecha),DATA6008$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Ordinario") +xlab("") +
ylab("")

a9<- ggplot(DATA6009,
aes(as.Date(DATA6009$Fecha),DATA6009$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Sumario") +xlab("") +
ylab("")

a10<- ggplot(DATA6010,
aes(as.Date(DATA6010$Fecha),DATA6010$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Verbal Sumario")
+xlab("") + ylab("")

a11<- ggplot(DATA6011,
aes(as.Date(DATA6011$Fecha),DATA6011$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Recusación") +xlab("") +
ylab("")

a12<- ggplot(DATA6012,
aes(as.Date(DATA6012$Fecha),DATA6012$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Solicitudes") +xlab("") +
ylab("")

a14<- ggplot(DATA6014,
aes(as.Date(DATA6014$Fecha),DATA6014$NumDia,group="60")) + geom_line()
+
  scale_x_date(labels = date_format("%m-%d"))+ggtitle("Med. protección")
+xlab("") + ylab("")
```

```
vplayout <- function(x, y) viewport(layout.pos.row = x, layout.pos.col = y)

grid.newpage()
pushViewport(viewport(layout = grid.layout(4, 3)))

print(a1, vp = vplayout(1,1))
print(a2, vp = vplayout(1,2))
print(a3, vp = vplayout(1,3))
print(a4, vp = vplayout(2,1))
print(a5, vp = vplayout(2,2))
print(a6, vp = vplayout(2,3))
print(a7, vp = vplayout(3,1))
print(a8, vp = vplayout(3,2))
print(a9, vp = vplayout(3,3))

print(a10, vp = vplayout(4,1))
print(a11, vp = vplayout(4,2))
print(a14, vp = vplayout(4,3))
```

b. Script en R para el modelo de regresión lineal de juicios resueltos por mes.

```
df <- read.table("analisis_obj1_total_mensual_ta_svm.csv",header = TRUE,sep =
";")

plot(df$Mes, df$Cantidad,pch=1,col=c('black'), xlab = "Meses", ylab = "Cantidad de
casos resueltos", main = "Regresión lineal de juicios \n resueltos por mes " )

# Creamos el modelos d regresion lineal
model = lm(df$Cantidad ~ df$Mes, df )

# agregamos la linea de tendencia
abline(model,lty = "dotted", col = "red")

summary(model)
predictedY <- predict(model, df)
points(df$Mes, predictedY, col = "blue", pch=4)
#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}

#calculando el MAE
mae <- function(error){
  mean(abs(error))
```

```
}

#calculando rmse del svm
errorRMSE <- model$residuals
svrPredictionRMSE <- rmse(errorRMSE)
svrPredictionRMSE

#calculando el mae del svm
error <- df$Cantidad - predictedY
svrPredictionMAE <- mae(error)
svrPredictionMAE
```

c. Script en R para el modelo de regresión no lineal aplicando el algoritmo SVM para los juicios resueltos por mes.

```
library(ggplot2)
library(scales)
library(grid)
df <- read.table("analisis_obj1_total_mensual_ta_svm.csv",header = TRUE,sep =
";")

plot(df$Mes, df$Cantidad,pch=1,col=c('black'), xlab = "Meses", ylab = "Cantidad de
casos resueltos", main = "Regresión no lineal de juicios \n resueltos por mes " )
lines(df$Mes, df$Cantidad)

#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}

#calculando el MAE
mae <- function(error){
  mean(abs(error))
}

#aplicando el algoritmo svm al modelo de regresión
model <- svm(df$Cantidad ~ df$Mes, df)
predictedY <- predict(model, df$Mes)
points(df$Mes, predictedY, col = "green", pch=4)
lines(df$Mes,predictedY,col = "red")

summary(model)
```

```
error <- df$Cantidad - predictedY
svrPredictionRMSE <- rmse(error)
svrPredictionRMSE
#calculando el mae del svm
svrPredictionMAE <- mae(error)
svrPredictionMAE
```

d. Script en R para el modelo de regresión lineal para el promedio de resolución de juicios por mes.

```
library(e1071)

df <- read.table("analisis_obj1_promedio_general.csv",header = TRUE,sep = ";")

plot(df$Mes, df$Promedio,pch=1,col=1 , type="l",xlab = "Meses", ylab =
"Tiempo(días)", main = "Regresión Lineal del tiempo promedio \n de los tipos de
acción relacionados" )

# Creamos el modelos d regresion lineal
model = lm(df$Promedio ~ df$Mes, df )

# agregamos la linea de tendencia
abline(model,lty = "dotted", col = "red")

summary(model)
predictedY <- predict(model, df)
points(df$Mes, predictedY, col = "blue", pch=4)

#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}
error <- model$residuals  # same as data$Y - predictedY
predictionRMSE <- rmse(error)
predictionRMSE
#calculando el MAE
mae <- function(error){
  mean(abs(error))
}

predictionRMSE <- mae(error)
predictionRMSE
```

e. Script en R para el modelo de regresión no lineal utilizando SVM para el promedio de resolución de juicios por mes.

```
library(e1071)
df <- read.table("analisis_obj1_promedio_general.csv",header = TRUE,sep = ";")

plot(df$Mes, df$Promedio,pch=1,col=1 , type="l",xlab = "Meses", ylab =
"Tiempo(días)", main = "Regresión no lineal del tiempo promedio \n de los tipos de
acción relacionados" )

#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}

#calculando el MAE
mae <- function(error){
  mean(abs(error))
}

#aplicando el algoritmo svm al modelo de regresión lineal para observar la diferencia
de ambos modelos.

model <- svm(df$Promedio ~ df$Mes, df)
predictedY <- predict(model, df$Mes)


points(df$Mes, predictedY, col = "green", pch=4)
lines(df$Mes,predictedY,col = "red")

summary(model)
#calculando rmse del svm
error <- df$Promedio - predictedY
svrPredictionRMSE <- rmse(error)
svrPredictionRMSE
#calculando el mae del svm
svrPredictionMAE <- mae(error)
svrPredictionMAE
```

Anexo 7: scripts del objetivo 2

a. Script en postgresql para obtener los datos para el análisis

```
SELECT  date_part('month',  "FechaResolucion")  as mes,round(avg("NumDias"),2)
as promedio FROM analisis_judicial
 GROUP BY mes
 order by mes asc
```

b. Script en R para la regresión lineal

```
df <- read.table("analisis_objetivo2_promedio_tiempo_mes.csv",header = TRUE,sep
= ";")

plot(df$Mes, df$Promedio,pch=1,col=1 , type="l",xlab = "Meses",
     ylab = "Tiempo(días)", main = "Regresión lineal del tiempo vs meses \n de la
materia familia niñez y adolescencia" )

# Creamos el modelos d regresion lineal
model = lm(df$Promedio ~ df$Mes, df )

# agregamos la linea de tendencia
abline(model,lty = "dotted", col = "red")
summary(model)
# se predice los nuevos valores
predictedY <- predict(model, df)
points(df$Mes, predictedY, col = "blue", pch=4)

#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}

#calculando el MAE
mae <- function(error){
  mean(abs(error))
}

# error RMSE
error <- model$residuals  # same as data$Y - predictedY
predictionRMSE <- rmse(error)
predictionRMSE

predictionMAE <- mae(error)
predictionMAE
```

c. Script en R para la regresión no lineal utilizando SVM

```
df <- read.table("analisis_objetivo2_promedio_tiempo_mes.csv",header = TRUE,sep = ";")

plot(df$Mes, df$Promedio,pch=1,col=1 , type="l",xlab = "Meses",
     ylab = "Tiempo(días)", main = "Regresión no lineal del tiempo promedio para resolución de juicios en \n la materia familia niñez y adolescencia" )

#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}

#calculando el MAE
mae <- function(error){
  mean(abs(error))
}

#aplicando el agoritmo svm al modelo de regresión lineal para observar la diferencia de ambos modelos.

model <- svm(df$Promedio ~ df$Mes, df)
predictedY <- predict(model, df$Mes)

points(df$Mes, predictedY, col = "green", pch=4)
lines(df$Mes,predictedY,col = "red")

summary(model)
error <- df$Promedio - predictedY
svrPredictionRMSE <- rmse(error)
svrPredictionRMSE
#calculando el mae del svm
svrPredictionMAE <- mae(error)
svrPredictionMAE
```

Anexo 8: scripts de objetivo 3

a. Script en postgresql para obtener los datos a ser analizados

```
SELECT  date_part('month',  "FechaResolucion")  as mes, count("NumDias") as promedio FROM analisis_judicial
 GROUP BY mes
 order by mes asc
```

b. Script en R para el modelo de regresión lineal

```
df <- read.table("analisis_obj3_promedio_general.csv",header = TRUE,sep = ";")

plot(df$Mes, df$Cantidad,pch=1,col=1 , type="l",xlab = "Meses",
     ylab = "Casos resueltos", main = "Regresión lineal de casos resueltos respecto a
meses en \n la materia familia niñez y adolescencia" )

# Creamos el modelos d regresion lineal
model = lm(df$Cantidad ~ df$Mes, df )

# agregamos la linea de tendencia
abline(model,lty = "dotted", col = "red")
summary(model)

# se predice los nuevos valores
predictedY <- predict(model, df)
points(df$Mes, predictedY, col = "blue", pch=4)

#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}

#calculando el MAE
mae <- function(error){
  mean(abs(error))
}

# error RMSE
error <- model$residuals  # same as data$Y - predictedY
predictionRMSE <- rmse(error)
predictionRMSE

predictionMAE <- mae(error)
predictionMAE
```

c. Script en R para el modelo de regresión no lineal aplicando el modelo SVM

```
df <- read.table("analisis_obj3_promedio_general.csv",header = TRUE,sep = ";")

plot(df$Mes, df$Cantidad,pch=1,col=1 , type="l",xlab = "Meses",
     ylab = "Casos resueltos", main = "Regresión no lineal de casos resueltos respecto
a meses en \n la materia familia niñez y adolescencia" )

#Calculando el error RMSE
rmse <- function(error)
{
  sqrt(mean(error^2))
}

#calculando el MAE
mae <- function(error){
  mean(abs(error))
}

#aplicando el agoritmo svm al modelo de regresión lineal para observar la diferencia
de ambos modelos.

model <- svm(df$Cantidad ~ df$Mes, df)
predictedY <- predict(model, df$Mes)

points(df$Mes, predictedY, col = "green", pch=4)
lines(df$Mes,predictedY,col = "red")

summary(model)
error <- df$Cantidad - predictedY
svrPredictionRMSE <- rmse(error)
svrPredictionRMSE
#calculando el mae del svm
svrPredictionMAE <- mae(error)
svrPredictionMAE
```