
**Universidad CENFOTEC**

**Maestría en Tecnologías de Bases de Datos**

**Documento final de Proyecto de Investigación Aplicada 2**

# Modelo de minería de datos para la predicción de delitos menores en Costa Rica

**Karen Ovares Garro**

**Patrick Robinson Martínez**

**Febrero, 2019**

## Tabla de Contenido

## Lista de figuras

## Lista de tablas

## Resumen ejecutivo

Este proyecto buscaba la identificación, creación e implementación de un modelo predictivo de minería de datos que permitiera pronosticar la probabilidad de si un delito menor se puede llevar a cabo o no en un determinado lugar de Costa Rica, basado en datos históricos de denuncias ante el Organismo de Investigación Judicial de Costa Rica. Estos datos son las estadísticas policiales del Poder Judicial, las cuales son publicadas en su sitio web como parte de su política organizacional de Justicia Abierta.

Este modelo fue desarrollado utilizando la metodología de CRISP-DM, basada en seis fases que facilitaron el estudio de los datos y la implementación de un modelo exitoso.

**Palabras clave:** Modelo de minería de datos, delito menor, predicción, CRISP-DM, Poder Judicial, Organismo de Investigación Judicial, Costa Rica, regresión, clasificación.

# Capítulo 1. Introducción

## 1.1 Generalidades

Uno de los principales ejes de trabajo del Poder Judicial durante el 2018 fue la seguridad ciudadana, evitando la comisión de delitos contra las personas y protegiendo sus bienes. El Gobierno de la República de Costa Rica ha destinado varios programas de desarrollo para la prevención de delitos y ha impulsado reformas de ley, así como el aumento en el número de oficiales de seguridad para promover la convivencia ciudadana y la seguridad pública.

## 1.2 Antecedentes del problema

El Poder Judicial cumple con la función de garantizar la obtención de una justicia de calidad, asegurando la convivencia y el desarrollo pacíficos, junto con la colaboración de los ciudadanos y de otras entidades de bien público; sin embargo, la inseguridad ciudadana ha venido creciendo a lo largo de los últimos años acarreando efectos negativos para la sociedad. Durante el 2015, los datos del Organismo de Investigación Judicial (OIJ) reportaron que una persona es asaltada cada 40 minutos. La percepción de la población (según una encuesta de la firma CID Gallup en el 2018) reveló que un 42% considera que existe un aumento en la delincuencia y un 51% asegura que ésta se mantiene igual.

En el 2015, el foco de prioridad del Gobierno de la República de Costa Rica impulsó al Poder Judicial a formar una plataforma de recolección y digitalización de datos de las denuncias con el objetivo de fortalecer las bases de un Gobierno Judicial Abierto y el Poder Judicial asumió el compromiso de crear una política de apertura de los datos.

Hasta este punto, la implementación de esta plataforma tecnológica permitía al Poder Judicial dar un mejor servicio a la ciudadanía en general y modernizar el accionar contra la inseguridad ciudadana. Después de recolectar los datos de estadísticas policiales durante tres años y medio, surgió la oportunidad de desarrollar técnicas de minería de datos y modelos predictivos donde se

pudieran analizar los datos y producir una retroalimentación sobre posibles delitos antes de que estos sean cometidos.

### 1.3 Definición y descripción del problema

*"Desde el 2011 el OIJ recibe un promedio de cerca de 80,000 casos al año. El 85% de ellos son denuncias contra ignorado, es decir, situaciones en que las víctimas no pueden identificar al autor del presunto delito".* - Estado de la Justicia 2017.

Sólo en el 2019 en promedio se dio un asalto por hora en el país, de acuerdo con los datos estadísticos registrados por el OIJ. Debido a este incremento en la tasa de delitos en Costa Rica, y a la falta de recursos del Poder Judicial destinados a la investigación de las denuncias, surgió la necesidad de impulsar una iniciativa que pudiera facilitar la resolución del problema de raíz.

Tal iniciativa consistió en el análisis de los datos abiertos del Poder Judicial sobre delitos menores, para generar un modelo predictivo por medio de técnicas de minería de datos que permitieron predecir la probabilidad de que ocurra un delito en un determinado lugar al momento en que una persona esté cerca de la zona o planee transitar a una hora específica, de forma que se pudiera contribuir sustancialmente con la seguridad ciudadana al reducir el índice de delitos menores en Costa Rica.

### 1.4 Justificación

El Poder Judicial mantiene una base de estadísticas policiales, la cual es actualizada mensualmente con reportes de asaltos, hurtos y robos. Conforme se analizaron los datos recolectados de estas incidencias, se pudo obtener retroalimentación sobre eventos de delitos menores en Costa Rica y pasar a ser una importante fuente de información para los organismos involucrados. De ahí que las técnicas de minería de datos aplicadas en este proyecto pretendían abastecer al Poder Judicial con las herramientas necesarias para combatir la criminalidad en el país y al mismo tiempo servir como mecanismo informativo a los ciudadanos.

Entre las principales ventajas de utilizar la minería de datos mediante la aplicación de modelos probabilísticos resalta la obtención de información que no se podría conseguir a simple vista. A pesar de que los datos son concretos, los modelos matemáticos permitieron predecir la eventualidad de futuros incidentes. Estos modelos fueron probados y presentaron un índice de confiabilidad tal, que contribuyeron a la toma de decisiones tácticas y estratégicas.

Este proyecto tuvo como iniciativa el dotar al Poder Judicial de una herramienta de análisis para la toma de decisiones estratégicas que le permitieran orientar los focos de atención de prevención en las zonas más propensas a la práctica de delitos, y proveer a los ciudadanos una plataforma de información que les ayudara a identificar situaciones de riesgo en el momento preciso.

## 1.5 Viabilidad

### 1.5.1 Punto de vista técnico

El Poder Judicial cuenta con las herramientas para la recolección de información sobre los delitos por medio de las denuncias presentadas ante el Organismo de Investigación Judicial (OIJ). Estos mismos datos han sido publicados en el sitio web del Poder Judicial, los cuales son clasificados como datos abiertos y de fácil acceso para cualquier ente que esté interesado en utilizarlos.

### 1.5.2 Punto de vista operativo

Para la realización de este proyecto se utilizaron datos públicos de denuncias interpuestas ante el Poder Judicial y publicados en su sitio web como estadísticas policiales. Estos datos fueron depurados y clasificados para mostrar la frecuencia y el tipo de víctima de los diferentes delitos ocurridos en Costa Rica. Tales datos han sido actualizados mensualmente desde el año 2015 hasta la actualidad. Desde el punto de vista operativo, el funcionamiento de la organización no se vio afectado o interrumpido durante la investigación.

### 1.5.3 Punto de vista económico

Para la realización de este proyecto no fue necesario realizar una inversión monetaria, ya que el modelo no fue vendido al Poder Judicial. Adicionalmente se utilizó RStudio como aplicación para la creación del modelo de minería de datos y la misma es gratis, por lo cual no se necesitó invertir en este ámbito. Sin embargo, cabe aclarar que el costo promedio por hora de un especialista en minería de datos con suficiente experiencia para realizar esta investigación ronda los $15, a su vez que el salario para este mismo puesto inicia desde los $3,000 mensuales.

## 1.6 Objetivos

Para la identificación de los objetivos se seleccionó la taxonomía de Bloom, gracias a que su estructura ordena jerárquicamente los procesos cognitivos, lo cual facilitó el desarrollo del proyecto.

### 1.6.1 Objetivo general

Diseñar un modelo de minería de datos para la predicción de delitos menores en Costa Rica.

### 1.6.2 Objetivos específicos

- Definir la pregunta que busca ser respondida con el modelo.
- Comprender cuáles son los datos necesarios, cómo se recolectarán y su significado preliminar.
- Preparar los datos para ser consumidos por el modelo.
- Proponer un algoritmo que se ajuste al problema que se quiere resolver.
- Evaluar los resultados del modelo.
- Desplegar el modelo para que pueda seguir recibiendo datos actualizados y logre cumplir con su función.

## 1.7 Alcances y limitaciones

### 1.7.1 Alcances

El proyecto buscaba entregar un documento escrito donde se definiera el proceso de desarrollo, investigación y evaluación del mismo, cumpliendo con los objetivos generales y específicos planteados al inicio de la investigación.

El ciclo de vida del proyecto estuvo basado en el modelo de CRISP-DM (Cross Industry Standard Process for Data Mining) como estándar en el análisis de datos. El proyecto buscaba procesar y analizar la información para determinar las métricas de confiabilidad, precisión y utilidad del modelo de minería escogido.

Los datos disponibles hasta el momento de la implementación abarcaban los delitos menores cometidos desde el año 2015 hasta el mes de setiembre de 2018. Estos datos contenían el tipo de delito, el sub-delito, fecha del suceso, descripción de la víctima, descripción de la sub-víctima, edad de la víctima, género de la víctima, nacionalidad de la víctima, provincia, cantón y distrito.

### 1.7.2 Limitaciones

La información que alimentó el modelo de minería de datos se obtuvo del Poder Judicial quien queda libre de responsabilidad ante cualquier proceso de recolección y sumarización de los datos.

Los resultados arrojados por el modelo dependieron estrictamente de los datos y no se implementaron validaciones adicionales.

La limitación inicial que presentó la construcción del modelo fue la de la falta de registros en el conjunto de datos que indicaran cuándo no se había cometido un delito, ya que solo se contaba con los registros que correspondían a delitos que sí se llevaron a cabo. Ante esta faltante se procedió a crear registros generando todas las posibles combinaciones que no existieran en el conjunto de datos original y clasificándolas como 0 (no fue víctima).

Una vez que se contaba con un conjunto de datos completo y con dos clasificaciones para la variable dependiente “EsVictima” (0 y 1) surgió una nueva limitación: esta vez se trataba de la cantidad de registros que se tenían en total,

alrededor de mil seiscientos millones, por lo cual se seleccionó únicamente el distrito de Hospital de la provincia de San José, para un total de tres millones de filas aproximadamente, puesto que es el distrito con más delitos reportados.

## 1.8 Marco de referencia organizacional y socioeconómico

### 1.8.1 Historia

El 15 de setiembre de 1821 los costarricenses constituyeron un gobierno propio, y en diciembre de ese mismo año se formuló el Pacto de Concordia, el cual fue considerado como el primer documento constitucional de Costa Rica. En dicho pacto se creó un tribunal para administrar la justicia y éste fue el primer cimiento de la Corte Suprema.

El 24 de setiembre de 1824 se dispuso la división del Estado en tres poderes: Ejecutivo, Legislativo y Judicial. En la rama judicial, el poder podía residir en una Corte Suprema de Justicia, pero hasta el 25 de enero de 1825 se creó constitucionalmente el Poder Judicial con la Ley Fundamental del Estado Libre de Costa Rica, atribuyendo su ejercicio a una Corte Superior de Justicia que se llegó a instalar hasta el 1° de octubre de 1826.

En 1871 se redactó una nueva Carta Fundamental, donde se estableció que el Poder Judicial quedaría conformado por la Corte Suprema de Justicia y demás tribunales y juzgados que la ley estableciera.

A partir del 29 de marzo de 1887, con la Ley Orgánica de Tribunales, se estableció por primera vez la independencia del Poder Judicial.

El 7 de noviembre de 1949 se emitió la Constitución que rige al país hasta la fecha y a partir de ese momento se dan una serie de innovaciones en la organización del Poder Judicial.

En mayo de 1957 se estableció la asignación al Poder Judicial de una suma no menor del seis por ciento de los ingresos ordinarios calculados para el año

económico, hecho de gran importancia por cuanto le garantiza al Poder Judicial un mínimo de ingresos, sin importar cuál sea el gobierno de turno.

Con la Ley 7128 del 18 de agosto de 1989 y el número 7135 del 11 de octubre del mismo año y la Ley de la Jurisdicción Constitucional, se logró uno de los más importantes cambios dentro del Sistema Judicial Costarricense al crearse la Sala Constitucional, con el poder de conocer las declaraciones de inconstitucionalidad de cualquier clase de normas jurídicas y de los actos sujetos al derecho público. Adicionalmente se le otorgaron dos funciones más: mayor jerarquía y la potestad de resolver conflictos entre los poderes del Estado, incluidos el Tribunal Supremo de Elecciones y las demás entidades u órganos que indique la Ley.

Para los años noventa el Poder Judicial se orientó a lograr una estructura más moderna, acorde a la realidad que debía enfrentar, a fin de poder cumplir con su deber. Durante los últimos años se han promovido procesos de reformas que han conducido al Poder Judicial a su modernización para que pueda cumplir con la demanda de resolución de conflictos que le son planteados en los Tribunales de Justicia.

### 1.8.2 Tipo de negocio y mercado meta

En el marco referencial de este proyecto se define al Poder Judicial de Costa Rica como el principal órgano encargado de administrar la justicia y contribuir con el bienestar de la sociedad. Por ende, en función de fortalecer y asegurar las demandas de la población adecuadamente, este proyecto se enfocó en la prevalencia de las funciones del Poder Judicial y en la atención a las demandas por parte de la sociedad.

### 1.8.3 Misión, Visión y Valores

**Misión**

“Administrar justicia en forma pronta, cumplida, sin denegación y en estricta conformidad con el ordenamiento jurídico, que garantice calidad en la prestación de servicios para las personas usuarias que lo requieran”.

**Visión**

“Ser un Poder Judicial que garantice a la persona usuaria el acceso a la justicia y resuelva sus conflictos con modernos sistemas de organización y gestión; compuesto por personal orientado por valores institucionales compartidos, conscientes de su papel en el desarrollo de la nación y apoyado en socios estratégicos”.

**Valores**

- **Iniciativa:** Tenemos la capacidad de orientar la acción innovadora y creativa para hacer mejor nuestras funciones.
- **Integridad:** Actuamos con rectitud y transparencia.
- **Compromiso:** Actuamos con responsabilidad para cumplir nuestros fines.
- **Honradez:** Trabajamos correctamente conforme a las normas morales, diciendo la verdad y siendo personas justas.
- **Responsabilidad:** Cumplimos con los deberes, obligaciones y compromisos asumiendo las consecuencias de nuestros actos.
- **Excelencia:** Realizamos con alto desempeño todas las acciones.

### 1.8.4 Políticas institucionales

**Política de Justicia Abierta**

La Justicia Abierta es una forma de gestión pública aplicada al Poder Judicial que busca fortalecer la transparencia, hacer uso de un lenguaje más comprensible y facilitar el acceso a la información, de forma que las personas puedan tener acceso a información sobre el quehacer institucional, lo cual facilita la rendición de cuentas, el debate público y la participación ciudadana; esto proporciona que se generen espacios de encuentro y canales de comunicación con los ciudadanos para que estos asuman un rol activo en las políticas y propuestas del Poder Judicial.

La Justicia Abierta se sustenta en tres principios que orientan su implementación: transparencia, participación y colaboración. Estos principios están correlacionados y de cada uno de ellos deriva una serie de ejes, los cuales a su vez definen las acciones por seguir.

- **Transparencia:** Responsabilidad de garantizar el derecho de acceso y la comprensión de la información pública sin mayores limitaciones, rendir cuentas sobre su gestión y propiciar la integridad.
    - **Acceso a la información pública:** Es el derecho que tienen los ciudadanos de acceder y comprender información pública y el deber del Poder Judicial de proporcionar tal información de manera oportuna.
    - **Apertura de datos:** Publicación de datos libres de controles y conforme a los estándares internacionales. La publicación de la información institucional debe ser consistente y perdurable, según los requerimientos de los datos abiertos.
    - **Rendición de cuentas:** Es el deber que tiene el personal judicial de responder por sus actos, por el cumplimiento de deberes y funciones, por el uso de recursos y fondos públicos.

- **Integridad, probidad y anticorrupción:** Mecanismos orientados al buen gobierno y a la lucha contra acciones que lesionen los valores, principios y recursos del Poder Judicial.

- **Participación:** Proceso democrático que garantiza la contribución responsable, activa y sostenida de la población en el diseño, la toma de decisiones y la ejecución de políticas del Poder Judicial.
  - **Interacción y diálogo:** La población puede comunicar las demandas sobre los servicios, exigir sus derechos y atención a sus necesidades y esperar una respuesta oportuna por parte de la institución.
  - **Seguimiento, control y evaluación ciudadanos:** Los pobladores pueden establecer una vigilancia sobre políticas, programas, proyectos, planes y procesos ejecutados en el Poder Judicial, o bien, conocer su impacto.
  - **Incidencia:** Incluir a la ciudadanía en espacios de toma de decisiones.

- **Colaboración:** Involucrar a la población en el diseño, ejecución y evaluación de políticas, programas, proyectos, planes y otras acciones propias del Poder Judicial.
  - **Alianzas:** Son acuerdos que se establecen entre el Poder Judicial y otras organizaciones públicas o privadas, así como también sociedad civil, para concretar vínculos de cooperación y emprender acciones conjuntas.
  - **Co-creación:** Desarrollo de procesos conjuntos entre el Poder Judicial y otros actores sociales para el diseño, gestión, ejecución y evaluación de políticas, programas, proyectos, planes y otras acciones.
  - **Redes de trabajo y apoyo:** Espacios en los cuales el Poder Judicial participa junto con la población para planificar, coordinar, construir, atender y dar seguimiento a temáticas relacionadas con el sistema de administración de justicia.

## 1.9 Estado de la cuestión

En este apartado se muestra el proceso de la revisión sistemática que se llevó a cabo, tomando como base la plantilla publicada en el libro Systematic Review in Software Engineering (Biolchini et.al; 2007).

### 1.9.1 Planificación de la revisión

#### 1.9.1.1 Formulación de la pregunta

Reconocer los diferentes modelos de minería de datos que hayan sido utilizados para la predicción de delitos menores.

Foco de la pregunta

En esta revisión sistemática se pretendía identificar las iniciativas o estudios realizados en torno a modelos de minería de datos para la predicción de delitos menores.

Amplitud y calidad de la pregunta

- **Problema:** La inseguridad ciudadana es uno de los principales factores que amenazan la convivencia pacífica en una sociedad y contribuyen a la generación de conductas violentas en las calles. A pesar de la labor de los entes de justicia, no existe una forma sistemática ni automatizada de combatir los delitos de manera proactiva; de ello resulta la inquietud por mejorar el análisis de los datos históricos para aplicar modelos de minería de datos para la predicción de posibles delitos menores.
- **Pregunta de investigación:** ¿Cuáles modelos de minería de datos se adaptan mejor y presentan un alto índice de confiabilidad al análisis predictivo de delitos menores en Costa Rica con los datos abiertos del Poder Judicial?
- **Palabras clave y sinónimos:**
    - Minería de datos: Data Mining, Exploración de Datos.

- Delitos: crimen, policía o análisis predictivo.

- **Intervención:** Predicción acertada de delitos menores.
- **Control:** En esta revisión sistemática se contó con los datos abiertos del Poder Judicial con el histórico de delitos menores cometidos en Costa Rica durante los últimos 3 años.
- **Resultado:** El resultado de esta revisión fue la identificación de un modelo de minería de datos que se ajustara de manera confiable a la predicción de delitos menores en Costa Rica, con datos abiertos del Poder Judicial.
- **Medida de salida:** Índice de confiabilidad del modelo de minería de datos en el contexto de entrenamiento y prueba del modelo.
- **Población:** Publicaciones sobre la aplicación de modelos de minería de datos en el contexto de predicción de delitos menores.
- **Aplicación:** El beneficiario directo de esta revisión sistemática fue el Poder Judicial y consecuentemente la sociedad en materia de seguridad ciudadana.
- **Diseño experimental:** Ningún meta-análisis fue aplicado.

#### 1.9.1.2 Selección de fuentes

Definición del criterio de selección de fuentes

El criterio de selección de las fuentes estuvo estrictamente ligado a la confiabilidad de los expertos en la materia, así como la accesibilidad web usando motores de búsqueda con palabras clave y sinónimos especificados en esta revisión.

Lenguajes de estudio

Los estudios primarios fueron en inglés.

Identificación de fuentes

**Método de selección de fuentes:** Investigación por medio de consultas en buscadores web y compañías de publicaciones en línea.

**Lista de fuentes:**

- Data Mining and Knowledge Discovery
- Scholar Google
- ACM Digital Library
- IEEE Xplore

**Cadenas de búsqueda:** (“Data Mining” or “Analysis”) AND “Crime” AND “Predictive”

Selección de fuentes después de la evaluación

Todas las fuentes cumplieron con el criterio de calidad.

Comprobación de las fuentes

Todas las fuentes fueron aprobadas.

1.9.1.3 Selección de los estudios

Definición de estudios

**Definición del criterio de inclusión y exclusión de estudios:** Los estudios debían presentar modelos de minería de datos aplicados a la predicción de delitos en el contexto de seguridad ciudadana, por lo cual, el análisis del contenido de las publicaciones por medio del índice y del resumen ejecutivo con base a las palabras clave dio la relevancia para la revisión sistemática.

**Procedimiento para la selección de los estudios:** El procedimiento inició con la aplicación de la cadena de búsqueda en las fuentes seleccionadas. Posterior a eso se evaluó el contenido por medio de la relevancia del índice y del resumen ejecutivo con base a las palabras clave. Finalmente, se refinó la selección con una lectura más profunda de las publicaciones seleccionadas.

**Definición de tipos de estudio:** Se seleccionaron los tipos de estudios que cumplieran con el criterio de selección de fuente y que pasaran el procedimiento definido para la selección del estudio.

### 1.9.2 Ejecución de la revisión (y evaluación de la ejecución)

#### 1.9.2.1 Selección de la ejecución

Los siguientes estudios fueron seleccionados.

| Fuente | Estudio | Autor | Año |
|---|---|---|---|
| Data Mining and Knowledge Discovery | Decision support systems for police: Lessons from the application of data mining techniques to "soft" forensic evidence | Giles Oatley, Brian Ewart, John Zeleznikow | 2006 |
| Data Mining and Knowledge Discovery | Crime Data Mining | Shyam Varan Nath | 2007 |
| Data Mining and Knowledge Discovery | Predictive Policing | Craig D. Uchida | 2014 |
| Data Mining and Knowledge Discovery | Social Network Analysis in Predictive Policing | Mohammad A. Tayebi, Uwe Glässer | 2016 |
| ACM | Analyzing and Predicting Spatial Crime Distribution Using Crowdsourced and Open Data | Alexandros Belesiotis, George Papadakis, Dimitrios Skoutas | 2018 |
| IEEE Xplore | Crime Pattern Detection Using Data Mining | Shyam Varan Nath | 2006 |
| IEEE Xplore | Crime Pattern Detection, Analysis & Prediction | Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav | 2017 |
| IEEE Xplore | Crime Analysis and Prediction Using Data Mining | Shiju Sathyadevan, Surya Gangadhar | 2014 |
| Google Scholar | An Enhanced Algorithm to Predict a Future Crime using Data Mining | Malathi. A, Santhosh Baboo | 2011 |
| Google Scholar | Survey of Data Mining Techniques on Crime Data Analysis | Revatthy Krishnamurthy, J. Satheesh Kumar | 2012 |

Tabla 1: Selección de estudios iniciales. Información obtenida de diferentes fuentes.

#### 1.9.2.2 Extracción de la información

**Inclusión de la información y exclusión del criterio de definición**

La información extraída de los estudios debía contener técnicas, conceptos básicos, historia, casos o cualquier otra iniciativa que presentara la aplicación de la minería de datos para la predicción de crímenes.

## Formularios de extracción de datos

| |
|---|
| **Extracción de resultados objetivos:** <Estos detalles se pueden obtener de cada estudio> |
| **Identificación del estudio:** <Incluye título, autores o año de publicación> |
| **Metodología del estudio:** <Método utilizado para conducir el estudio> |
| **Resultados del estudio:** <Efectos obtenidos por medio de la ejecución del estudio> |
| **Problemas del estudio:** <Limitaciones del estudio> |
| **Extracción de resultados subjetivos:** <Estos detalles son apreciaciones personales> |
| **Información mediante autores:** <Información obtenida directamente del autor> . |
| **Impresiones generales y abstracciones:** <Información de otros lectores> |

Tabla 2: Formulario de extracción de datos. Fuente: Elaboración propia.

## Extracción de la ejecución

En esta sección se muestra la extracción de los datos e información relevante de cada uno de los estudios seleccionados.

| **Extracción de resultados objetivos** |
|---|
| **Identificación del estudio:** Oatley, G., Ewart, B. & Zeleznikow, J. Decision support systems for police: Lessons from the application of data mining techniques to "soft" forensic evidence (2006). |
| **Metodología del estudio:** Los autores trabajan en torno a la aplicación de minería de datos para la detección y prevención de crímenes trabajando de la mano con 3 servicios policiales. El apartado busca referenciar la evidencia forense junto con los datos de tiempo y ubicación de los delitos para crear un patrón de conducta que pueda ser procesado por un sistema computacional existente y automatizar así el análisis de dichos datos. |
| **Resultados del estudio:** La necesidad de un enfoque multidisciplinario, en el contexto de criminalística, para el diseño de un sistema computacional que pueda automatizar la predicción de delitos. |
| **Problemas del estudio:** La orientación del estudio está dirigida a una base de datos específica y a un sistema computacional existente. |
| **Extracción de resultados subjetivos** |
| **Información mediante autores:** No fue solicitada. |
| **Impresiones generales y abstracciones:** El estudio muestra cualidades esenciales para el desarrollo de un sistema de procesamiento de datos policiales para la prevención y detección temprana de delitos. |

Tabla 3: Primera fuente literaria investigada. Información obtenida de Oatley, G; B. & Zeleznikow, J (2006).

| Extracción de resultados objetivos |
|---|
| **Identificación del estudio:** Nath S.V. Crime Data Mining (2007) |
| **Metodología del estudio:** El autor busca capturar los años de experiencia humana por medio de la Minería de Datos, para ello aplica diferentes técnicas como algoritmos de clustering para detectar patrones y acelerar la resolución de crímenes. También utiliza aprendizaje semi-supervisado para incrementar la exactitud de las predicciones. Finalmente, se utilizan gráficos geo-espaciales para ayudar a los agentes policiales a actuar fácilmente. |
| **Resultados del estudio:** Se obtuvieron resultados significativos en los atributos para la predicción de delitos, ayudando a los agentes policiales a detectar crímenes y realizar identificación de sospechosos. |
| **Problemas del estudio:** N/A |
| **Extracción de resultados subjetivos** |
| **Información mediante autores:** No fue solicitada. |
| **Impresiones generales y abstracciones:** El estudio demuestra una gran posibilidad de aportar una estructura adecuada para la aplicación de técnicas de minería de datos en el contexto de seguridad ciudadana para entidades de seguridad. |

Tabla 4: Segunda fuente literaria investigada. Información obtenida de

Nath S. V. Crime Data Mining (2007).

| Extracción de resultados objetivos |
|---|
| **Identificación del estudio:** Uchida C.D. Predictive Policing (2014) |
| **Metodología del estudio:** El autor busca describir el origen de la predicción de delitos y los conceptos principales asociados a la minería de datos y Análisis Predictivo Espacial. También elabora en los principales retos alrededor de esta tendencia. |
| **Resultados del estudio:** Definición de los principales conceptos de predicción de delitos en el contexto de minería de datos y análisis geoespacial. |
| **Problemas del estudio:** No existen referencias técnicas sobre la elaboración de un modelo de minería de datos. |
| **Extracción de resultados subjetivos** |
| **Información mediante autores:** No fue solicitada. |
| **Impresiones generales y abstracciones:** El estudio brinda los conceptos clave y el trasfondo de la necesidad de aplicación de técnicas de minería de datos en la detección de delitos antes de su comisión. |

Tabla 5: Tercera fuente literaria investigada. Información obtenida de Uchida C.D. (2014).

| Extracción de resultados objetivos |
| --- |
| **Identificación del estudio:** Shyam Varan Nath. Crime Pattern Detection Using Data Mining (2006). |
| **Metodología del estudio:** Para este estudio se usa el algoritmo de k-means con algunos arreglos para ayudar en el proceso de reconocer los patrones de los crímenes. Se aplicaron estas técnicas a datos de crímenes reales. También se utilizó una técnica de aprendizaje supervisado para descubrimiento de conocimiento y para ayudar a incrementar la precisión de la predicción. |
| **Resultados del estudio:** La implementación fue sencilla y puede trabajar con graficación geoespacial que ayuda a mejorar la productividad de los detectives. |
| **Problemas del estudio:** Puede ayudar a los detectives, pero no puede reemplazarlos. La minería de datos es sensible a la calidad de datos. Mapear datos reales a atributos de minería de datos no es siempre una tarea fácil y casi siempre se requiere de algún experto en materia de crímenes. |
| **Extracción de resultados subjetivos** |
| **Información mediante autores:** No fue solicitada. |
| **Impresiones generales y abstracciones:** El uso de k-means para la creación de patrones resultó interesante para el tipo de investigación ya que muestra tendencias que pueden resultar útiles para las investigaciones. |

Tabla 6: Cuarta fuente literaria investigada. Información obtenida de Shyam Varan Nath (2014).

| Extracción de resultados objetivos |
| --- |
| **Identificación del estudio:** Shiju Sathyadevan, Devan M.S, Surya Gangadharan. Crime Analysis and Prediction Using Data Mining (2014) |
| **Metodología del estudio:** Los autores buscan predecir las regiones que tienen alta probabilidad de ocurrencia de crímenes. Inicialmente extraen los datos históricos desde una fuente no estructurada, seguidamente utilizan Naive Bayes para clasificar los datos. El tercer paso es la selección del patrón utilizando el algoritmo Apriori. El cuarto paso se basa en la predicción utilizando el modelo de árbol de decisión, y por último la visualización de los resultados. |
| **Resultados del estudio:** La utilización de Bayes resultó un 90% de precisión. El programa logra predecir el crimen por regiones de la india en un día en particular. |
| **Problemas del estudio:** Sería más preciso si se considerara el estado o la región, así como también considerar la hora en que sucede el evento. Se deberían incluir otros factores para así poder brindar mejores resultados. |
| **Extracción de resultados subjetivos** |
| **Información mediante autores:** No fue solicitada. |
| **Impresiones generales y abstracciones:** El estudio es bastante completo y muestra que se siguió un proceso adecuado para llegar a la predicción, así como también muestra el resultado de la aplicación de Bayes y de Apriori y cómo esto contribuye de buena manera en el estudio. |

Tabla 7: Quinta fuente literaria investigada. Información obtenida de Shiju Sathyadevan, Devan M.S, Surya Gangadharan. (2014).

**Resolución de divergencias entre los revisores:** No existe divergencia alguna.

### 1.9.3 Análisis de resultados

**Resultados cálculo estadístico:** No se utilizó ningún cálculo estadístico.

**Presentación de resultados**

| Fuente | Estudios | Relevantes | Excluidos | Primarios |
|---|---|---|---|---|
| Data Mining and Knowledge Discovery | 15 | 4 | 11 | 3 |
| Scholar Google | 10 | 3 | 7 | 0 |
| ACM Digital Library | 4 | 1 | 3 | 0 |
| IEEE Xplore | 8 | 2 | 6 | 2 |

Tabla 8: Resultados de fuentes literarias. Fuente: Elaboración propia.

**Análisis de sensibilidad:** No aplicable.

- **Gráficos:** No aplicable.

**Comentarios finales**

- **Número de estudios:** 37 estudios encontrados, 5 seleccionados.
- **Sesgo de búsqueda, selección y extracción:** No fue definido.
- **Sesgo de publicación:** No fue definido.
- **Variación entre revisores:** No hay variación alguna.
- **Aplicación de resultados:** Los estudios mostraron formas de aplicar modelos de minería en la predicción de crímenes.
- **Recomendaciones:** Ninguna.

## Capítulo 2. Marco teórico o conceptual

En este apartado se describe el marco conceptual de las generalidades expuestas en el capítulo anterior. Este marco se considera necesario para interpretar el enfoque aplicativo de este proyecto para el diseño de un modelo de minería de datos para la predicción de delitos menores en Costa Rica.

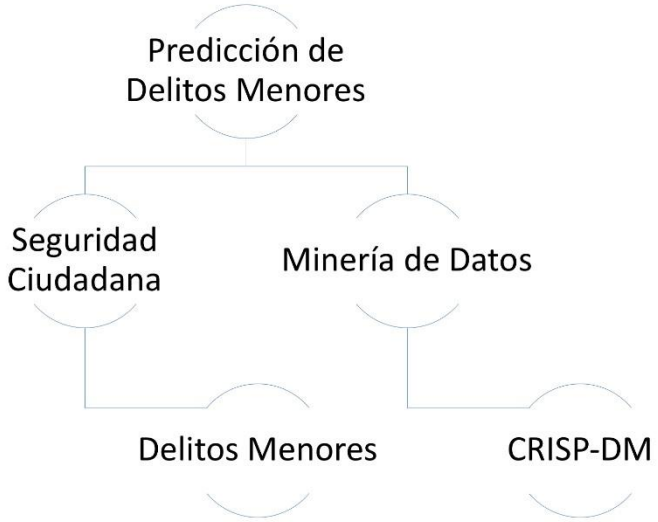


Figura 1: Relación entre seguridad ciudadana y minería de datos. Fuente: Elaboración propia.

### 2.1 Seguridad ciudadana

La seguridad ciudadana se considera como un conjunto de acciones por parte del Estado, en colaboración con la ciudadanía y organizaciones, tanto públicas como privadas, en pro de la seguridad de sus habitantes y de sus bienes. Generalmente se busca marcar un espacio de convivencia pacífica erradicando la violencia y la comisión de delitos.

Cada una de estas acciones se ajustan al derecho de los habitantes y al marco democrático de cada país, buscando la armonía con el ejercicio de los derechos humanos.

El objetivo de la seguridad ciudadana es garantizar una libertad concreta protegiendo la vida, la integridad y el patrimonio de los habitantes.

## 2.2 Delitos menores

*"Conforme a lo que hoy plantea la dogmática es una conducta típica (acción u omisión), antijurídica y culpable. Sus elementos son entonces la tipicidad, la antijuridicidad y la culpabilidad. Se trata de una definición tripartita del delito; la tipicidad, adecuación de un hecho determinado con la descripción que de él hace un tipo legal; la antijuridicidad, la contravención de ese hecho típico con todo el ordenamiento jurídico, y la culpabilidad, el reproche porque el autor pudo actuar de otro modo, es decir, conforme al orden jurídico".* - Bustos Ramírez, Manual de Derecho Penal 3era. Edición (1984). Editorial Ariel S.A., Barcelona.

Se describe típicamente como una infracción al derecho penal que es penalizada por la ley. Dicha conducta es contraria al ordenamiento jurídico y arremete contra la seguridad de los ciudadanos y de sus bienes.

Según el esquema de trabajo planteado, se definieron los siguientes tipos de delitos para este proyecto:

- **Robo** - Delito que se comete apoderándose con ánimo de lucro de una cosa mueble ajena, empleando violencia o intimidación sobre las personas, o fuerza en las cosas.
- **Hurto** - Delito consistente en tomar con ánimo de lucro cosas muebles ajenas contra la voluntad de su dueño, sin que concurran las circunstancias que caracterizan el delito de robo.
- **Asalto** - Acometer repentinamente y por sorpresa.
- **Robo de vehículo** - Específico a vehículos.
- **Tacha de vehículo** - Consiste en robar o hurtar objetos extrayéndolos de un vehículo.

## 2.3 Minería de datos

La minería de datos es un grupo de técnicas aplicadas con tecnologías de información para explorar un conjunto de datos de manera automatizada para

encontrar patrones, tendencias o reglas que logren describir el comportamiento de dichos datos. Todo esto mediante prácticas estadísticas y algoritmos que ayudan a comprender el contenido de un repositorio de datos y realizar predicciones en un contexto determinado.

## 2.4 CRISP - DM

CRISP-DM (del inglés Cross Industry Standard Process for Data Mining) es un proceso modelado para la descripción de metodologías de minería de datos a utilizar.

CRISP-DM divide el proceso de minería de datos en seis fases principales:

- **Comprensión del negocio** - Se aplica el entendimiento tanto de los objetivos como de los requerimientos del proyecto, desde un punto de vista general.
- **Comprensión de los datos** - Se ejecuta la recolección de los datos para su entendimiento e identificación de información oculta.
- **Preparación de los datos** - Se preparan los datos para construir un conjunto limpio de datos para las herramientas de modelado.
- **Modelado** - Se ejecutan varias técnicas de modelado y se calibran los parámetros para obtener así los mejores resultados.
- **Evaluación** - Desde el punto de vista analítico se evalúa la calidad del modelo.
- **Explotación** - En esta fase se puede presentar un reporte o se puede implementar un proceso a nivel organizacional, todo dependiendo de los requerimientos.

Este modelo no es lineal, por lo cual simboliza el ciclo natural de la minería de datos, permitiendo moverse de una fase a otra tanto como sea necesario para alcanzar los objetivos.

# Capítulo 3. Marco metodológico

En el presente capítulo se presenta la metodología que se empleó en esta investigación.

## 3.1 Tipo de investigación

En este proyecto se utilizó una investigación de tipo aplicada, ya que las características de este tipo de investigación concuerdan con lo que se buscaba. Se pretendía predecir si se daría un delito menor en una región específica de Costa Rica con ayuda de técnicas de minería de datos. En Costa Rica no existía un sistema informático que pudiera predecir delitos, por lo que esta iniciativa buscaba ayudar a las autoridades a mejorar la seguridad del país.

"*A la investigación aplicada también se le denomina activa o dinámica y se encuentra íntimamente ligada a la básica ya que depende de sus descubrimientos y aportes teóricos. Aquí se aplica la investigación a problemas concretos, en circunstancias y características concretas. Esta forma de investigación se dirige a una utilización inmediata y no al desarrollo de teorías*". (Ernesto A. Rodríguez Moguel 2005, pág. 23).

## 3.2 Alcance investigativo

El alcance investigativo de este proyecto fue exploratorio, ya que de acuerdo a la literatura encontrada se han hecho investigaciones de aplicaciones de minería de datos en la detección de crímenes, pero ninguna de ellas ha sido aplicada en Costa Rica.

"*Los estudios exploratorios se realizan cuando el objetivo es examinar un tema o problema de investigación poco estudiado, del cual se tienen muchas dudas o no se ha abordado antes. Es decir, cuando la revisión de la literatura reveló que tan sólo hay guías no investigadas e ideas vagamente relacionadas con el problema de estudio, o bien, si deseamos indagar sobre temas y áreas*

*desde nuevas perspectivas*". (Roberto Hernández Sampieri, Carlos Fernández Collado, María del Pilar Baptista Lucio, 2010, pág. 79).

### 3.3 Enfoque

El enfoque aplicado en este proyecto fue un enfoque mixto, es decir, tanto cuantitativo como cualitativo. El primer enfoque surgió a manera de presentar los resultados estadísticos y la comparación entre las métricas de confiabilidad del modelo de minería de datos. El segundo enfoque exploró la viabilidad de utilizar modelos de minería de datos para la predicción de delitos menores.

### 3.4 Diseño

Pese a que el enfoque utilizado es mixto, el diseño buscaba hallar las relaciones entre las variables del conjunto de datos y proponer un modelo de minería que se ajustara a las cualidades del requerimiento de predicción de delitos menores. Para ello, fue necesaria la recolección de datos abiertos del Poder Judicial para el análisis inicial. Seguidamente, se aplicaron técnicas de minería de datos con el fin de concluir con un modelo que presentara las métricas lo suficientemente confiables para abordar en una propuesta formal. Dado todo esto, se obtuvo un parámetro de calificación que pudo exponer de manera efectiva la utilización de modelos de minería de datos para la predicción de delitos menores en Costa Rica.

### 3.5 Población y muestreo

El conjunto de datos utilizados correspondió a los reportes de delitos menores registrados por el OIJ a lo largo de las diferentes provincias del territorio costarricense desde el año 2015 hasta setiembre del 2018. Esto cubría la mayor parte de la población.

### 3.6 Instrumentos de recolección de datos

Los datos utilizados están a disposición en el sitio web de datos abiertos del Poder Judicial, donde están categorizados y segmentados por años y a la vez disponibles para descarga en diferentes formatos.

### 3.7 Técnicas de análisis de la información

A continuación se muestra un gráfico con el flujo de la técnica de análisis que se utilizó (CRISP-DM):

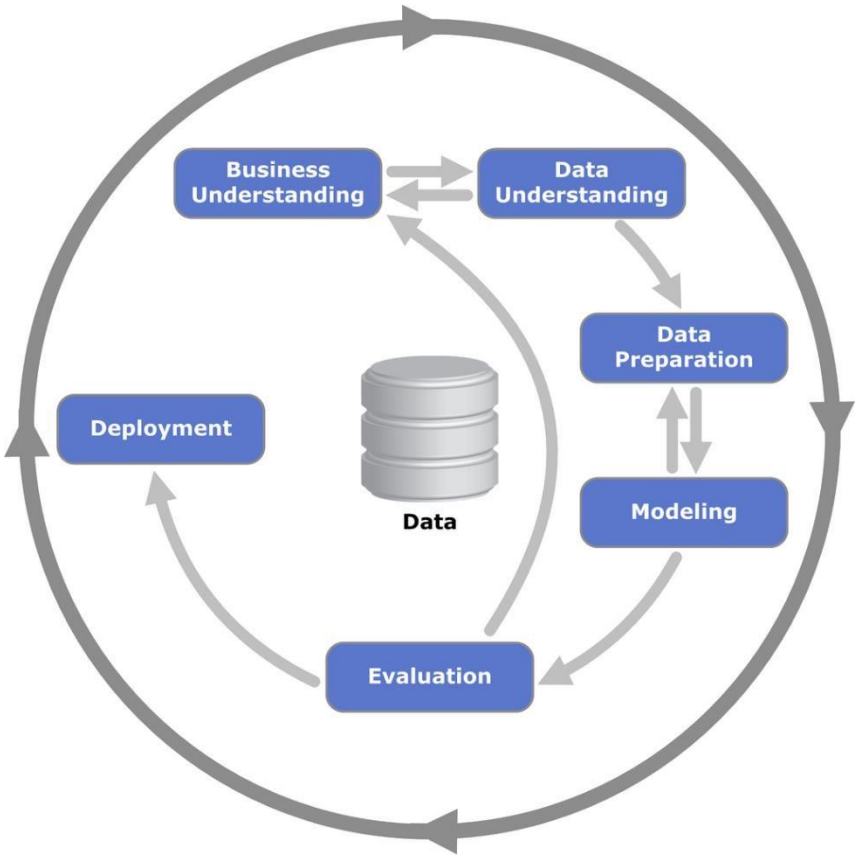


Figura 2: Diagrama CRISP-DM. Fuente: Elaboración propia (obtenida de CRISP-DM: La metodología para poner orden en los proyectos de Data Science).

1. **Entendimiento del negocio:** En este paso se indica que se debe entender el contexto del negocio para determinar cuál es realmente el problema a resolver, establecer los objetivos claros y además identificar cuáles son los datos que se van a necesitar.

2. **Entendimiento de los datos:** Abarca todas las actividades relacionadas con la construcción del conjunto de datos. Se recolectan los datos iniciales, se describen, se exploran y se verifica la calidad, de forma que en esta sección se pueda responder: ¿se tienen los datos adecuados para resolver el problema?

3. **Preparación de los datos:** Es la parte en donde se seleccionan los datos que son relevantes, se limpian, se crean algunos nuevos campos en caso de ser necesario y se les da el formato correcto, para que de esta manera se pueda construir el conjunto de datos.

4. **Modelado:** En este apartado se selecciona la técnica de modelado predictivo adecuada. Se crea un diseño de la prueba para probar la validez del modelo. Se utiliza un conjunto de datos de entrenamiento del cual se conocen los resultados y se le aplican algoritmos para garantizar que las variables en juego son realmente relevantes.

5. **Evaluación:** En esta fase se evalúa sobre todo si el modelo cumple con los objetivos del negocio, se determina si hay alguna razón de negocio que haga que el modelo no sea válido y de igual forma se puede probar integrándolo con las herramientas del negocio o con casos de la vida real.

6. **Explotación:** Se planea el despliegue del modelo, se documenta y se crea una estrategia. Se planea el monitoreo y el mantenimiento si el modelo va a ser parte de la operación diaria.  Finalmente se documenta que salió bien o qué salió mal.

# Capítulo 4. Análisis del diagnóstico

## 4.1 Entendimiento del negocio

Para comprender el negocio se necesitó seguir cada una de las tareas de las que consta la primera fase en el modelo de CRISP-DM, y así determinar los objetivos y requisitos del proyecto desde una perspectiva del negocio, para transformarlos inicialmente en objetivos técnicos y posteriormente en un plan de proyecto.

### 4.1.1 Objetivos del negocio

El objetivo principal para el negocio era poder realizar predicciones en cuanto a la probabilidad de ocurrencia de algún delito menor antes de que ocurriera. La fiabilidad de estas predicciones buscaba proporcionar un mejor servicio a la seguridad ciudadana a partir de los datos que se tenían.

#### 4.1.1.1 Contexto

Bajo el marco contextual, en el Poder Judicial se contó con una base de datos de delitos menores de los últimos 3 años, al igual que un mecanismo de recolección que hasta la fecha toma registro de dichos sucesos; sin embargo, no existía un estudio a profundidad sobre el comportamiento de estos delitos que pudiera llevar a conclusiones o patrones para realizar la predicción de futuros incidentes.

#### 4.1.1.2 Objetivos del negocio

En tanto a los objetivos del negocio, se intentó dar con la predicción de delitos menores de tal manera que se pudiera calcular una probabilidad del suceso fiable partiendo de los datos existentes y continuamente recolectados. Se podrían hacer muchas predicciones según el modelo de minería de datos aplicado; no obstante, en este proyecto se definieron los siguientes objetivos:

- Estimar la probabilidad de ocurrencia de un delito menor en un lugar determinado.
- Determinar los factores o atributos en una persona que aumentan la posibilidad de ser víctima de un delito menor.

Estos informes pueden ser de gran utilidad para las entidades del Poder Judicial a la hora de dirigir campañas de seguridad ciudadana o investigaciones judiciales, de tal forma que se atribuyen de mejor manera los recursos hacia los sectores más propensos a dichos delitos. Todo esto permitirá que el Poder Judicial mejore la calidad del servicio a la ciudadanía en cuanto a seguridad.

#### 4.1.1.3 Criterio de éxito

Desde la perspectiva del negocio, se estableció como criterio de éxito la realización de predicciones sobre el acontecimiento de delitos menores con un porcentaje de fiabilidad aceptable, de tal forma que se puedan dirigir campañas y esfuerzos de seguridad a focos de la población más propensos.

Al mismo tiempo, se consideró importante el poder concluir con un aprendizaje sobre la utilización de los datos en un modelo de minería de datos, más allá de si el modelo funciona o no.

### 4.1.2 Evaluación de la situación actual

Se contó con un conjunto de datos en archivos planos separados por coma, con información detallada de los delitos menores que se han cometido desde el 2015 hasta la actualidad, por lo que *a priori* se podía afirmar que se disponía de suficientes datos para poder realizar una evaluación técnica para la aplicación de un modelo de minería de datos.

**Inventario de recursos**

En cuanto a los recursos de software, se dispuso de los siguientes componentes:

- **Archivos de texto plano:** Cuatro conjuntos de datos en formato CSV con información relevante de los delitos menores acontecidos en los últimos 3 años. Dicha fuente de datos sirvió como base para la evaluación de diferentes modelos de minería de datos.
- **Software especializado:** Además de eso, se contó con herramientas de visualización y minería de datos como Microsoft Power BI, R Studio y SQL Server.

En cuanto a los recursos humanos, se disponía de varios departamentos de proveeduría del ámbito de seguridad ciudadana y judicial:

- **Ministerio de Seguridad Pública:** Del área de estructura interna se trabajó en conjunto con la Fuerza Pública y la reserva de fuerzas policiales para conocer los beneficios a impacto de la aplicación de predicciones en el tema de delitos menores. Cabe destacar que la dirección policial del Ministerio de Seguridad Pública es quien se encarga de ejecutar las acciones de seguridad ciudadana, de integridad territorial y el mantenimiento del orden público.
- **Poder Judicial de Costa Rica:** En esta área se trabajó mayormente con la dirección del Organismo de Investigación Judicial, encargado de realizar las investigaciones mismas y la recolección de los datos necesarios para su análisis, así como el Organismo de Digesto de Jurisprudencia que recopila, clasifica y publica los contenidos jurídicos de interés público.

Cabe destacar que ambos recursos trabajan en conjunto acorde al "*Convenio Marco de Colaboración Interinstitucional entre el Poder Judicial de la República de Costa Rica y el Ministerio de Seguridad Pública*".

**Requisitos, supuestos y restricciones**

Al no poder tener acceso al repositorio de datos central de manera directa, se limitó el modelo a permanecer bajo las restricciones de disponibilidad de los datos abiertos del Poder Judicial.

**Costos y beneficios**

Los datos que se utilizaron para este proyecto no supusieron ningún costo adicional al Poder Judicial, ya que los datos son de carácter público y están abiertos a los usuarios por descargas vía web.

Los beneficios, en tanto al carácter económico, no se podrían cuantificar directamente, dado que el objetivo principal era mejorar la calidad del servicio a la seguridad ciudadana y, por tanto, a la satisfacción de la opinión pública.

### 4.1.3 Objetivos de la minería de datos

Los objetivos en términos de minería de datos son:

- Predecir la probabilidad de que ocurra un delito menor en un momento determinado en función de los atributos expuestos en el conjunto de datos y con base a los patrones de ocurrencia histórica.
- Determinar los atributos/características de las personas más propensas a ser víctimas de un delito menor.
- Identificar los lugares más vulnerables a ser acometidos por delitos menores acorde a su historial de acontecimientos.

**Criterio de éxito de minería de datos**

Como criterio de éxito se estableció la determinación de un modelo de minería de datos que pudiera realizar un cálculo probabilístico sobre la sucesión de delitos menores en un momento y lugar determinados, con base en las características de una persona y con la utilización de datos históricos, el cual fue

dado por el algoritmo que se empleó a la hora de conseguir el modelo de minería de datos.

### 4.1.4 Plan de proyecto

Este proyecto se dividió en seis etapas para así organizar y estimar el tiempo de realización del mismo. Tales etapas se especificaron en el Capítulo 3.7 de este documento.

## 4.2 Entendimiento de los datos

En esta fase fue necesario adquirir los datos de los recursos disponibles para su propio entendimiento y verificar que fueran apropiados para las necesidades del proyecto. Una vez familiarizados los datos, se empezaron a formular las primeras hipótesis.

### 4.2.1 Recolección de datos

Los datos utilizados en este proyecto fueron referentes a las estadísticas policiales que incluyen información del hecho como: lugar, fecha, género, clasificación de la edad y nacionalidad de la víctima, así como el tipo de delito cometido. Estas estadísticas forman parte de los datos clasificados y depurados del Poder Judicial por medio de su portal de datos abiertos para mostrar la frecuencia y el tipo de víctima para cada delito que se puede materializar en las diferentes zonas del territorio nacional.

Para este proyecto se utilizó la distribución de archivos de texto separados por coma para el análisis:

- PJCROD_POLICIALES_V1-2015.csv
- PJCROD_POLICIALES_V1-2016.csv
- PJCROD_POLICIALES_V1-2017.csv
- PJCROD_POLICIALES_V1-2018.csv

### 4.2.2 Descripción de los datos

Los datos que se almacenaron en los archivos de texto separados por coma se encuentran de manera tabular y el detalle de los hechos están agrupados por años.

Estos archivos contienen la siguiente cantidad de registros:

| Nombre del documento | Registros |
|---|---|
| PJCROD_POLICIALES_V1-2015.csv | 54,925 |
| PJCROD_POLICIALES_V1-2016.csv | 58,535 |
| PJCROD_POLICIALES_V1-2017.csv | 59,335 |
| PJCROD_POLICIALES_V1-2018.csv | 50,501 |

Figura 3: Nombres de archivos y cantidad de registros. Fuente: Elaboración propia.

A continuación, se enlistan los datos adquiridos en cada archivo:

| Campo | Descripción | Tipo de dato | Es llave |
|---|---|---|---|
| Delito | Descripción del delito | Texto | No |
| Sub-delito | Descripción del sub-delito | Texto | No |
| Fecha | Fecha del delito | Fecha | No |
| Víctima | Descripción de la víctima | Texto | No |
| Sub-víctima | Descripción de la sub-víctima | Texto | No |
| Edad | Descripción de la edad | Texto | No |
| Género | Descripción del género de la víctima | Texto | No |
| Nacionalidad | Descripción de la nacionalidad de la víctima | Texto | No |
| Provincia | Descripción de la provincia | Texto | No |
| Cantón | Descripción del cantón | Texto | No |
| Distrito | Descripción del distrito | Texto | No |

Figura 4: Detalles de las columnas de los archivos. Fuente: Elaboración propia.

Estos datos contienen particularidades y patrones que a simple vista ayudaron a proporcionar las primeras hipótesis.

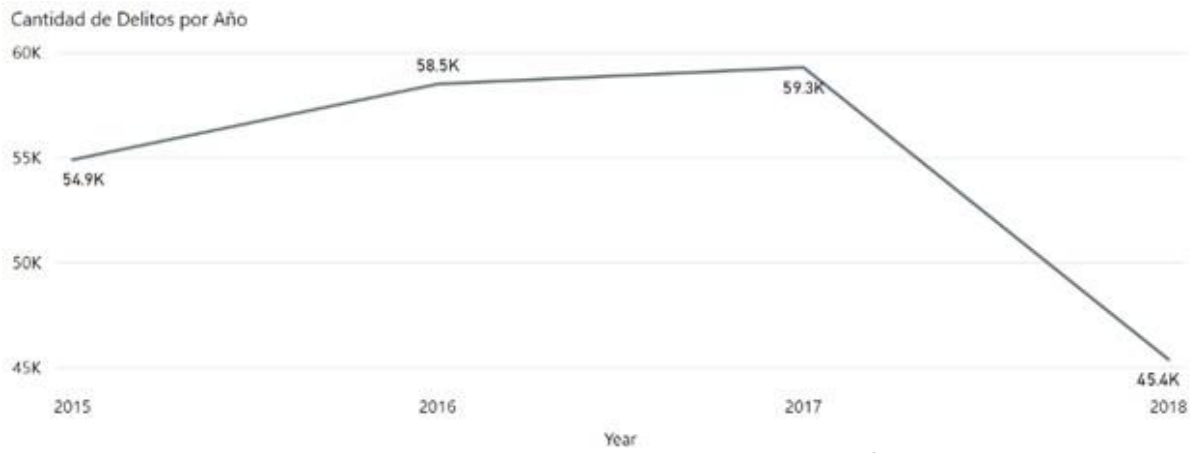

Figura 5: Cantidad de delitos por año. Fuente: Elaboración propia.

Del gráfico anterior se pudo inferir que en el año 2017 hubo más delitos reportados en comparación con el 2015 y 2016. En el 2018 se mostraron menos delitos reportados, ya que era el año en curso al momento de esta investigación y se contaba solo con datos de enero a setiembre.

### 4.2.3 Exploración de los datos

Con esta información se generaron análisis previos y así se logró sacar conclusiones preliminares que ayudaron a escoger el modelo indicado para la predicción de delitos menores.


Figura 6: Cantidad de delitos por víctima. Fuente: Elaboración propia.

Cabe destacar que el tipo de víctima con más delitos reportados fueron las personas físicas, con alrededor de noventa y cinco mil denuncias, seguido por delitos contra vehículos, vivienda y edificaciones, respectivamente. En este proyecto el enfoque principal fueron las personas físicas.

**Datos de las personas**

A continuación se muestran algunas visualizaciones en cuanto al perfil de las personas que fueron víctimas de delitos:

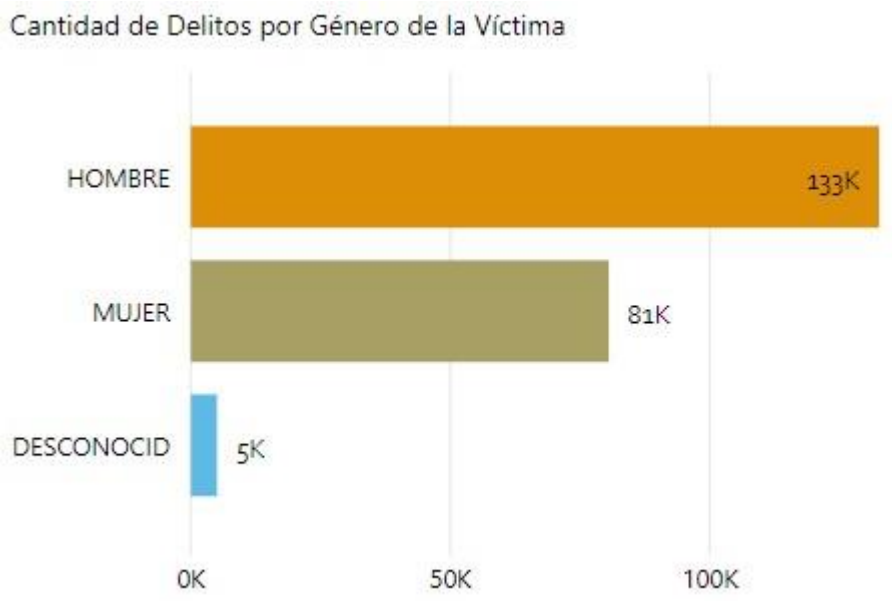


Figura 7: Cantidad de delitos por género de la víctima. Fuente: Elaboración propia.

La mayor cantidad de víctimas reportadas fueron hombres, con un 60,77% del total; 36.91% correspondió a mujeres, y el 2,32% fue desconocido.

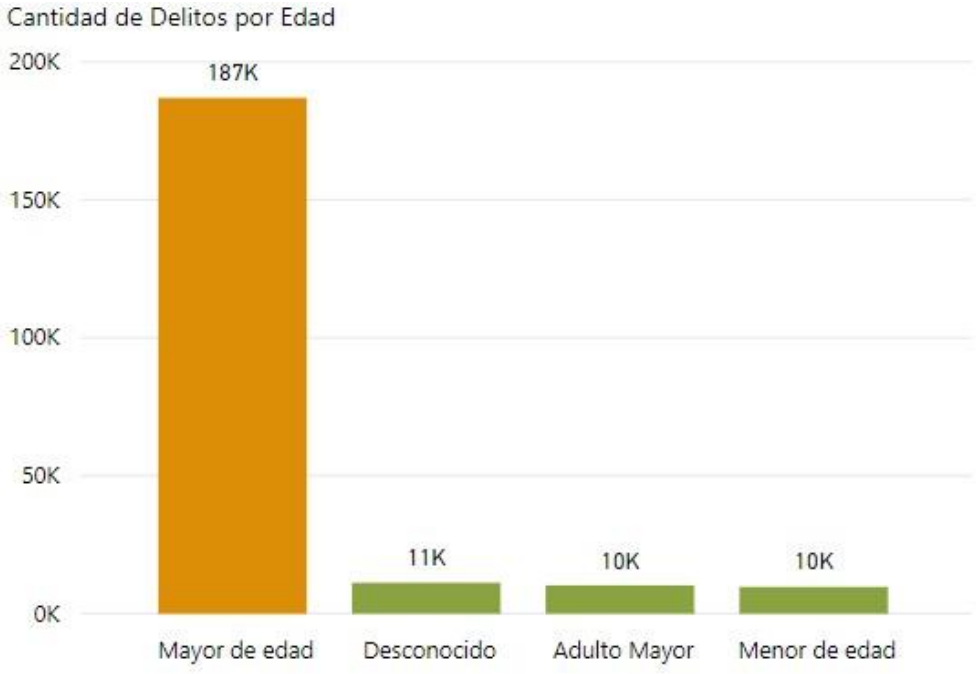


Figura 8: Cantidad de delitos por edad de la víctima. Fuente: Elaboración propia.

Alrededor de ciento ochenta y siete mil delitos reportados tenían como víctima a un mayor de edad, y esta fue la tendencia mayoritaria.

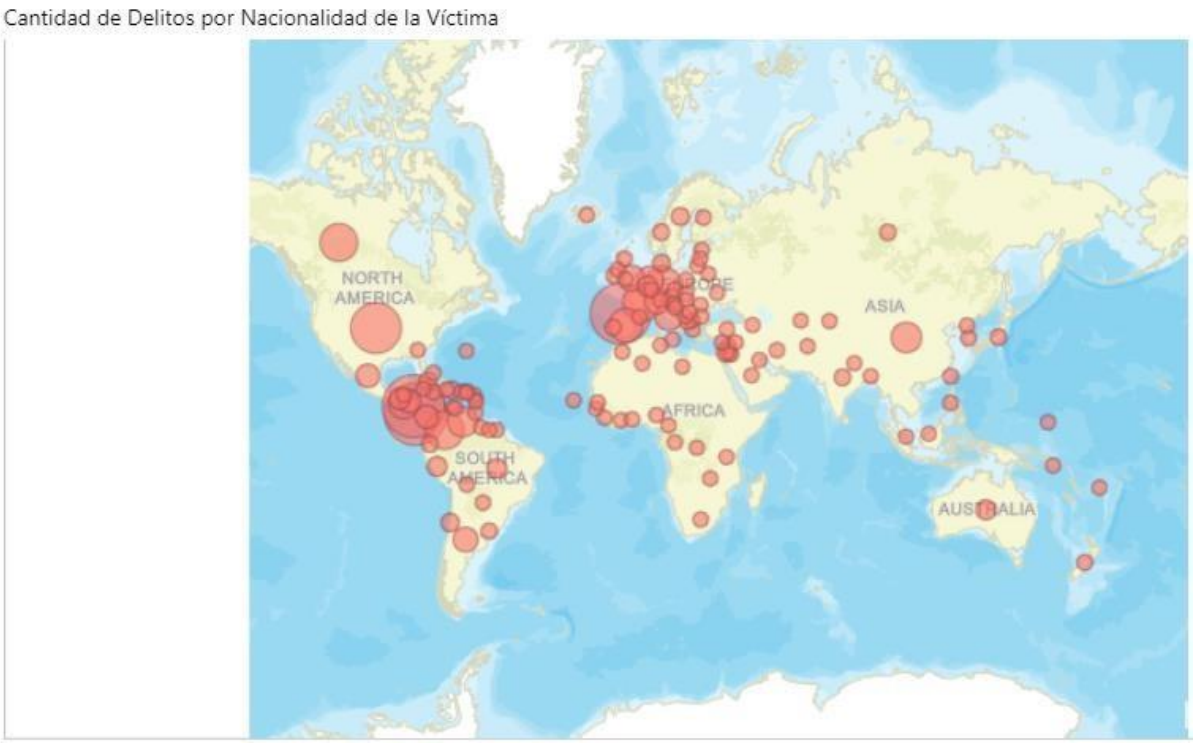


Figura 9: Cantidad de delitos por nacionalidad de la víctima en mapa. Fuente: Elaboración propia.

Las víctimas fueron personas de distintas nacionalidades, en su mayoría de nacionalidad costarricense; no obstante, también hubo víctimas de nacionalidad nicaragüense, estadounidense y colombiana, entre otras. A continuación, se muestra la distribución de registros agrupados por nacionalidad:

| Nacionalidad | CantDelitos |
|---|---|
| COSTA RICA | 172505 |
| NICARAGUA | 15778 |
| Desconocido | 13839 |
| ESTADOS UNIDOS | 3866 |
| COLOMBIA | 1099 |
| FRANCIA | 1070 |
| CANADA | 1043 |
| ALEMANIA | 953 |
| ESPAÑA | 818 |
| VENEZUELA | 625 |
| CHINA | 606 |
| EL SALVADOR | 578 |
| ITALIA | 492 |
| PAISES BAJOS (HOLANDA) | 402 |
| INGLATERRA | 371 |
| SUIZA | 336 |
| ARGENTINA | 332 |
| MEXICO | 284 |
| PANAMA | 283 |
| HONDURAS | 277 |
| GUATEMALA | 215 |
| **Total** | **218175** |

Figura 10: Cantidad de delitos por nacionalidad de la víctima. Fuente: Elaboración propia.

El gráfico anterior muestra la nacionalidad de las víctimas por cantidad de delitos. De acuerdo con los datos, las víctimas de nacionalidad costarricense fueron 172515, encabezando la lista, seguido por nicaragüenses, con 15778.

Así, esta información fue de gran utilidad para generar un perfil de persona que sea más propensa a sufrir un delito, por ejemplo, en la siguiente tabla se muestran dichas relaciones:

| CantDelitos | Edad | Genero | Nacionalidad |
|---|---|---|---|
| 93551 | Mayor de edad | HOMBRE | COSTA RICA |
| 57887 | Mayor de edad | MUJER | COSTA RICA |
| 8595 | Mayor de edad | HOMBRE | NICARAGUA |
| 6042 | Adulto Mayor | HOMBRE | COSTA RICA |
| 5624 | Mayor de edad | MUJER | NICARAGUA |
| 4952 | Menor de edad | HOMBRE | COSTA RICA |
| 4788 | Mayor de edad | HOMBRE | Desconocido |
| 3302 | Menor de edad | MUJER | COSTA RICA |
| 2945 | Desconocido | DESCONOCID | Desconocido |
| 2818 | Mayor de edad | MUJER | Desconocido |
| 2375 | Adulto Mayor | MUJER | COSTA RICA |
| 2176 | Desconocido | HOMBRE | COSTA RICA |
| 1623 | Mayor de edad | HOMBRE | ESTADOS UNIDOS |
| 1240 | Desconocido | HOMBRE | Desconocido |
| 1087 | Mayor de edad | MUJER | ESTADOS UNIDOS |
| 1063 | Desconocido | MUJER | COSTA RICA |
| 980 | Desconocido | DESCONOCID | COSTA RICA |
| 672 | Menor de edad | HOMBRE | Desconocido |
| 659 | Desconocido | MUJER | Desconocido |
| 606 | Mayor de edad | HOMBRE | COLOMBIA |
| 492 | Mayor de edad | HOMBRE | FRANCIA |
| 478 | Mayor de edad | HOMBRE | CANADA |
| 460 | Desconocido | HOMBRE | NICARAGUA |
| 417 | Adulto Mayor | HOMBRE | ESTADOS UNIDOS |
| 412 | Mayor de edad | HOMBRE | ESPAÑA |
| 404 | Mayor de edad | HOMBRE | ALEMANIA |
| 401 | Mayor de edad | MUJER | FRANCIA |
| 385 | Mayor de edad | MUJER | ALEMANIA |
| 359 | Mayor de edad | HOMBRE | VENEZUELA |
| **218175** | | | |

Figura 11: Cantidad de delitos por nacionalidad, edad y género de la víctima. Fuente: Elaboración propia.

En la tabla anterior se puede denotar que los datos mostraron un perfil de víctima de género masculino, mayor de edad y de nacionalidad costarricense como la que más delitos ha sufrido, con 93,551 eventos durante los últimos tres años.

## Datos del momento

Con la información de los delitos también se pudieron generar algunas visualizaciones sobre el comportamiento de tales delitos a través del tiempo.

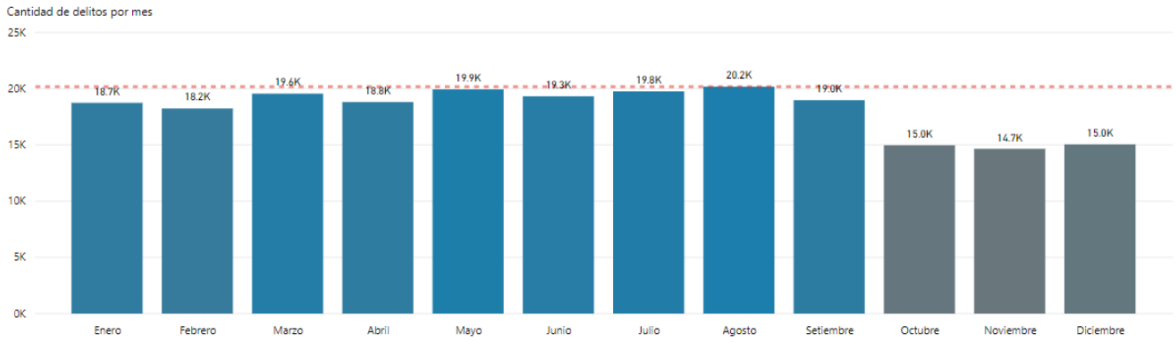


Figura 12: Cantidad de delitos por mes del año. Fuente: Elaboración propia.

El gráfico anterior muestra la cantidad de delitos por mes durante los años estudiados. Se debe tomar en cuenta que no se consideraron los datos del Poder Judicial para los meses de octubre 2018 hasta la actualidad (al momento de la realización de este proyecto), lo que pudo provocar la baja en estos últimos meses.

En el 2015, el mes con más denuncias fue julio, con cuatro mil novecientos. En el 2016 fueron julio y diciembre con cinco mil doscientos cada uno. En el 2017 se tuvieron cinco mil trescientos en marzo y noviembre.

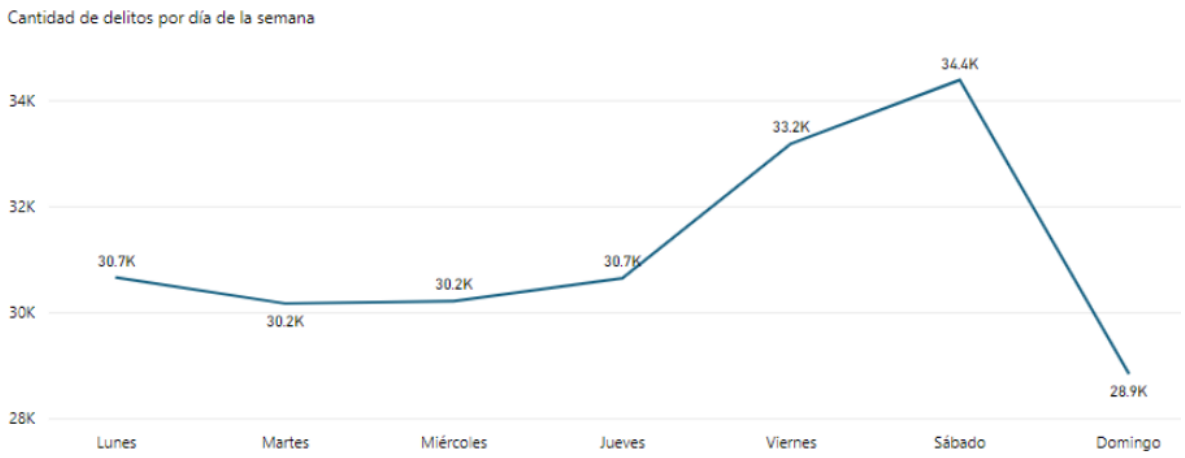


Figura 13: Cantidad de delitos por día de la semana. Fuente: Elaboración propia.

De acuerdo con los datos, el día más peligroso es el día sábado, ya que en este día ocurrieron treinta y cuatro mil cuatrocientos delitos.

## Datos del lugar

Por último, los datos también permitieron analizar el patrón de comportamiento de los delitos en las distintas zonas del país.

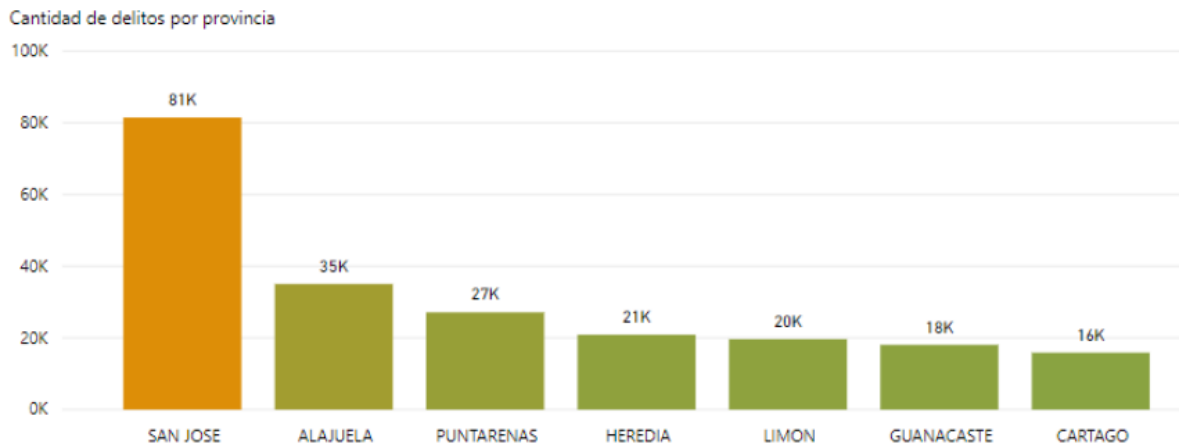


Figura 14: Cantidad de delitos por provincia. Fuente: Elaboración propia.

La provincia con más incidencia de delitos es San José, con alrededor de ochenta y un mil denuncias, seguido por Alajuela y Puntarenas.

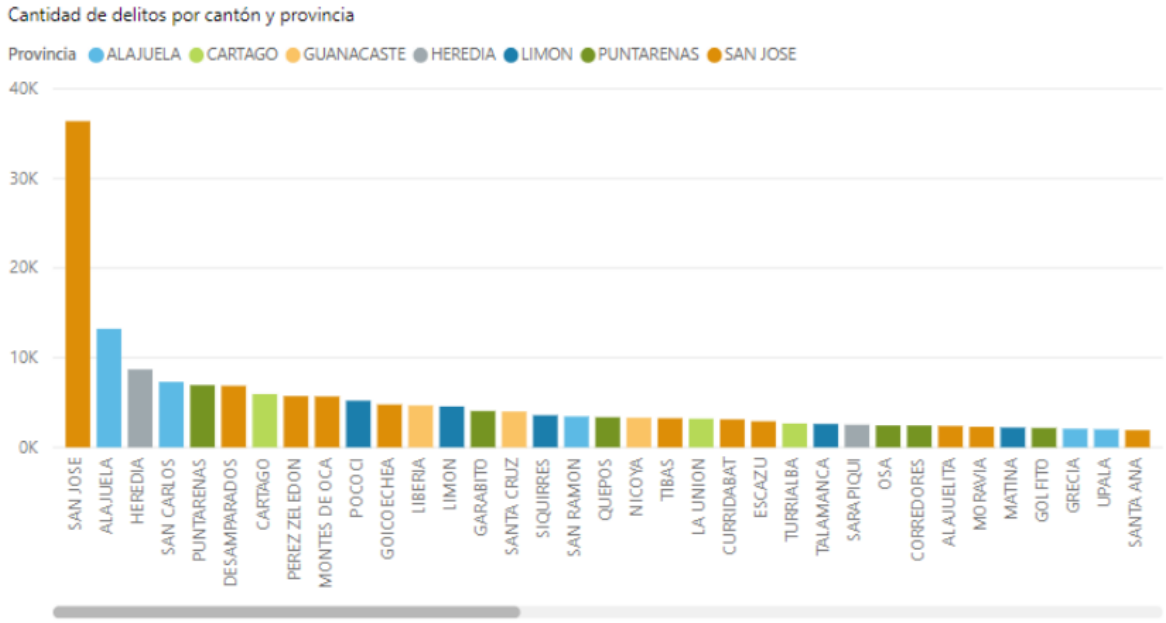

Figura 15: Cantidad de delitos por provincia y cantón. Fuente: Elaboración propia.

El gráfico anterior muestra los cantones con más delitos. El cantón de San José encabezó la lista, con 36 386 delitos, seguido por los cantones de Alajuela y Heredia. En los últimos lugares estuvieron el cantón de Dota de San José y Río Cuarto de Alajuela, con 129 y 98 delitos respectivamente.

### 4.2.4 Verificación de la calidad de datos

En cuanto al extracto del conjunto de datos como tal, se conocía de antemano que existían algunas limitaciones en el detalle de la hora en la que se acometieron los delitos y la edad de las víctimas. Sólo se contaba con el día de los hechos, pero si se hubiera tenido el dato de la hora, se hubiesen generado proyecciones acertadas en cuanto al momento específico en que podrían ocurrir delitos, así mismo, si se hubiera tenido la edad específica de las víctimas se habrían generado perfiles sobre las personas más propensas a tener un delito.

Sin embargo, aun con estas limitaciones se realizaron los análisis necesarios para identificar las aseveraciones más cercanas a la probabilidad de comisión de delitos en Costa Rica, es decir, con la información recopilada, y luego de realizar la exploración inicial, se afirmó que estos eran datos lo suficientemente completos como para cumplir los objetivos del proyecto.

Los datos contenían algunos errores de duplicidad en cuanto a los tipos de sub-delitos iguales, pero que se nombraban de manera diferente, como por ejemplo los delitos que fueron por descuido versus falta de cuidado.

Se descartó la posibilidad de algún riesgo de ruido en el proceso de minería de datos; tampoco se encontraron valores nulos, por lo que se aprovechó al máximo el conjunto de datos.

## 4.3 Preparación de los datos

En esta fase se intentó preparar los datos para adecuarlos a las técnicas de minería de datos que se utilizaron en este proyecto. Básicamente, la intención era seleccionar un subconjunto de datos, limpiarlos para mejorar su calidad, añadir nuevos datos de ser necesario y darles formato para utilizarlos en las herramientas de modelado que se definieron.

### 4.3.1 Selección de datos

En cuanto a los registros, se utilizaron los datos de todos los años disponibles en los archivos de texto plano distribuidos por el Poder Judicial en su portal de datos abiertos; sin embargo, dados los requerimientos del proyecto donde se requería saber la probabilidad de si un delito iba a suceder o no con base a una combinación de factores, el tipo de delito no fue requerido ni los atributos sub-derivados de éste. De la misma forma, tampoco se utilizó la sub-víctima puesto que para el alcance de este proyecto todavía no se demandaba definir un perfil tan específico de la persona, sino más bien estudiar la viabilidad del modelo con datos más superficiales.

Los atributos seleccionados para el análisis fueron:

1. Fecha: Texto, mil trescientos sesenta y ocho factores.
2. Víctima: Texto, cuatro factores.
3. Edad: Texto, tres factores.
4. Género: Texto, tres factores.
5. Nacionalidad: Texto, ciento treinta factores.
6. Provincia: Texto, siete factores.
7. Cantón:  Texto, ochenta y tres factores.
8. Distrito: Texto, trescientos ochenta y un factores.

La escogencia de estos atributos estuvo estrictamente ligada con la importancia que tienen tales atributos para suplir las necesidades de este proyecto y el cumplimiento de los objetivos planteados anteriormente.

Una vez hecha la primera selección de datos y realizada la construcción de los nuevos datos necesarios para la aplicación del modelo, se apreció que la cantidad de filas resultantes eran alrededor de mil seiscientos millones, por lo cual se decidió seleccionar únicamente el distrito de Hospital de la provincia de San José, para un total de tres millones de filas aproximadamente, ya que fue el distrito con más delitos reportados.

### 4.3.2 Integración de datos

Fue necesaria la integración de algunos conjuntos de datos para propiciar al modelo información completa. Bajo este escenario, se emplearon algunos métodos utilizados para combinar la información de múltiples conjuntos de datos.

**Datos combinados**

A pesar de que los datos disponibles en el portal del Poder Judicial están lo suficientemente completos, fue necesario combinar los datos en un solo conjunto

de datos para poder utilizarlos en el modelo. Se tuvo que combinar los datos de los diferentes años en un solo conjunto de datos que permitiera una manipulación más fácil. La combinación de datos se hizo utilizando SQL Server y luego fueron exportados en formato*.txt* para ser cargados a RStudio.

**Datos sumarizados**

La técnica de sumarización de los datos fue necesaria para sacar los posibles valores de cada atributo, los cuales fueron utilizados para obtener todas las posibles combinaciones entre todos los atributos; esto con el fin de conseguir las observaciones que se muestran en la sección de ´Registro generados´ de este documento.

### 4.3.3 Construcción de datos

A continuación, se incluyen algunas de las tareas que se llevaron a cabo para la construcción de datos derivados a partir de otros atributos, la transformación de atributos existentes o la creación de registros totalmente nuevos.

**Atributos derivados**

Se puede destacar la creación de atributos como de las columnas de MES y DÍA, puesto que son atributos derivados a partir de la FECHA. El motivo de la creación de estos atributos fue la necesidad de contar con más detalles sobre los eventos y de esta forma encontrar posibles patrones dentro del conjunto de datos.

Además de este atributo, se creó la columna "*EsVictima*" en su condición como valor binario para satisfacer así la necesidad de incluir en los distintos modelos una variable dependiente que pudiera ser utilizada en las predicciones.

**Registros generados**

Dado que el objetivo principal del proyecto era generar predicciones de probabilidad con base a situaciones donde sucede o no sucede un evento, se necesitaba contar tanto con registros que hicieran referencia a un delito cometido, como también registros referentes a delitos no cometidos. Gracias a que el Poder Judicial en su portal de Justicia Abierta solo cuenta con los registros de delitos que sí se llevaron a cabo, surgió la necesidad de crear registros nuevos. La metodología utilizada consistía en generar todas las posibles combinaciones de los valores de cada atributo que no existieran dentro del conjunto de datos y colocarles en la columna de “EsVictima” un 0. De esta forma se daba a entender que bajo esas condiciones no se dio un delito, ya que no formaba parte de los registros originales.

### 4.3.4 Limpieza de datos

A pesar de que el conjunto de datos contenía la información necesaria para poder cumplir con los objetivos de minería de datos expuestos en este proyecto, fue necesaria la limpieza de algunos atributos para poder así generar resultados más acertados. Al mismo tiempo algunas de las acciones tomadas sobre estos datos estaban enlazadas con los requerimientos de algunos modelos en cuanto a la utilización de caracteres especiales, entre otros.

**Caracteres especiales y símbolos**

Para generar algunos de los modelos fue necesario hacer una limpieza de caracteres especiales y símbolos que, durante el procesamiento de los datos, los modelos no lograron interpretar, por lo cual los valores “miércoles” y “sábado” del atributo DIA fueron modificados para eliminar las tildes. Por otro lado, algunas de las NACIONALIDADES contaban con paréntesis, guiones y comas que también generaron inconvenientes a la hora de cargar el conjunto de datos en el modelo.

**Valores desconocidos o nulos**

Por otra parte, las variables de edad, nacionalidad y género poseían valores descritos como “DESCONOCIDOS” para los cuales era necesaria la aplicación de un filtro de tal manera que se pudieran controlar los registros que no aportaban mayor valor a los objetivos de este proyecto.

## 4.4 Modelado

En esta sección se seleccionó la técnica de modelado predictivo adecuada. Se utilizó un conjunto de datos de entrenamiento del cual se conocían los resultados (si fueron o no víctimas de un delito) y se le aplicaron algoritmos para garantizar que las variables en juego eran realmente útiles para predecir la probabilidad de si el evento pudiera suceder o no.

### 4.4.1 Selección de la técnica de modelado

Las distintas técnicas de modelado utilizadas para la minería de datos permiten aplicar algoritmos supervisados, semi-supervisados y no-supervisados. Dado que se intentó asignar una clasificación a un segundo conjunto de datos de prueba a partir de un conjunto de datos ya clasificado (datos de entrenamiento), fue necesario aplicar un algoritmo de clasificación supervisada**.**

La clasificación trata de predecir la tendencia entre dos o más categorías, por lo que en este caso se hizo entre sí y no, representados por 1 y 0 respectivamente. Con base a esta sentencia, algunos modelos como la regresión logística, árboles de decisión, bosques aleatorios y las máquinas de soporte vectorial se adecuaban a los objetivos de este proyecto.

Independientemente del modelo a escoger, se partía del hecho de que cada registro formaba una clase de carácter dicotómico, esto es, solo dos resultados son posibles (1 o 0) y que para todos los atributos no se permitían valores perdidos o desconocidos.

También se consideró para todos los modelos un umbral de discriminación para el cálculo de la probabilidad de 0.5.

### 4.4.2 Generación del plan de prueba

Antes de construir el modelo fue necesario generar el procedimiento o mecanismo para probar su calidad y validez. Se construyeron conjuntos de prueba y entrenamiento como estructuras separadas para estimar la calidad del modelo, utilizando la función *"sample.split"*. Esta función ayudó a separar los datos en estos dos conjuntos en una relación predefinida mientras preservaba las proporciones relativas de diferentes etiquetas en el conjunto de datos. Para este proyecto se aplicó una proporción del 70% para el conjunto de datos de entrenamiento y un 30% para el conjunto de pruebas.

Es importante mencionar que esta técnica puede brindar diferentes resultados siempre que sea ejecutada, ya que no se asegura tener siempre la misma proporción de elementos positivos y negativos en cada uno de los conjuntos.

Adicionalmente para probar los modelos se utilizó una matriz de confusión y se calculó la sensibilidad, especificidad, exactitud y la tasa de error:

| | | **Predicción** | |
|---|---|---|---|
| | | **Positivos** | **Negativos** |
| **Observación** | **Positivos** | Verdaderos Positivos (VP) | Falsos Negativos (FN) |
| | **Negativos** | Falsos Positivos (FP) | Verdaderos Negativos (VN) |

Figura 16: Cantidad de delitos por provincia y cantón. Fuente: Data Mining and Knowledge Discovery.

- **VP** es la cantidad de *positivos* que fueron *clasificados correctamente* como positivos por el modelo.
- **VN** es la cantidad de *negativos* que fueron *clasificados correctamente* como negativos por el modelo.
- **FN** es la cantidad de *positivos* que fueron *clasificados incorrectamente* como negativos.
- **FP** es la cantidad de *negativos* que fueron *clasificados incorrectamente* como positivos.

**Exactitud** = (VP + VN) / Total

**Sensibilidad** = VP / Total Positivos

**Especificidad** = VN / Total Negativos

**Tasa de error** = (FP + FN) / Total

En los algoritmos de clasificación utilizados en la minería de datos supervisada es común utilizar matriz de confusión como medida de calidad en los conjuntos de prueba y entrenamiento.

### 4.4.3 Construcción del modelo

En esta sección se buscaba ejecutar las herramientas de modelado con los conjuntos de datos creados para tal propósito. Los modelos que mejor se ajustaban a los requerimientos de minería de datos fueron:

1. Regresión logística
2. Árboles de decisión
3. Máquinas de soporte vectorial
4. Bosques aleatorios

**Ajustes de parámetros**

Dependiendo de cada modelo, los parámetros a utilizar variaban, como por ejemplo en el caso de las máquinas de soporte vectorial, era necesario partir del hecho de que se necesitaba utilizar variables predictoras con dos o más niveles, por lo cual los atributos de 'provincia', 'cantón' y 'distrito' quedaron excluidos de los parámetros de este modelo, únicamente por las limitaciones expuestas en la sección ´Selección de datos´ de este documento.

Por otro lado, para poder usar el comando de Modelo Lineal Generalizado en R era necesario especificar que se debía modelar utilizando una distribución binomial, de tal forma que la variable de respuesta fuera categórica con dos posibles resultados. De la misma forma se utilizó el tipo de predicción *response* para obtener como resultado una etiqueta 0 o 1 en este caso.

```
reg.logistica <- glm(EsVictima ~ Mes + Dia + Victima + Edad + Genero + Nacionalidad,
                     family= 'binomial',
                     data = entrenamiento)

prediccion.log  <- predict(reg.logistica,newdata =prueba, type = 'response')
```

Figura 17: Regresión logística. Fuente: Elaboración propia.

En cuanto al árbol de decisión, el tipo de valor de retorno predictivo era preferiblemente una probabilidad, por lo que en su matriz las columnas que expusieron una probabilidad mayor al umbral de discriminación fueron las que definieron la categoría de la predicción.

```
modelo.arbol <- rpart(EsVictima ~ . , data = entrenamiento)

prediccion.arbol <- predict(modelo.arbol, newdata = prueba, type = 'prob')
```

Figura 18: Árbol de decisión. Fuente: Elaboración propia.

En el caso de las máquinas de soporte vectorial, de igual manera se utilizó un modelado de clasificación y un kernel radial, el cual era un enfoque adecuado ya que los datos no eran separables linealmente.

```
modelo.svm <- svm(EsVictima~ Mes + Dia + Victima + Edad + Genero + Nacionalidad
                  ,data = entrenamiento,
                  kernel="radial",
                  type='C-classification')

prediccion.svm <- predict(modelo.svm, newdata = prueba[-10])
```

Figura 19: Máquinas de soporte vectorial. Fuente: Elaboración propia.

Por último, se trabajó con un conjunto de los bosques aleatorios de 5 a 20 árboles.

```
modelo.RF<-randomForest(EsVictima ~
                         Mes + Dia + Victima + Edad +
                           Genero + Provincia + Canton + Distrito ,
                        data=entrenamiento,
                        ntree=5)

predicciones.RF<-predict(modelo.RF,newdata = prueba,type = 'prob')
```

Figura 20: Bosque aleatorio. Fuente: Elaboración propia.

## Modelos

Se ejecutaron los cuatro modelos sobre un conjunto de datos de entrenamiento del 70% y un 30% de pruebas. A continuación, los detalles del resultado de la herramienta en RStudio.

## 1. Regresión logística

Acorde a los parámetros empleados para la construcción del modelo, estos fueron los resultados iniciales, utilizando una primera versión de los datos aún con ciertos caracteres y registros que no habían sido filtrados:

```
> summary(reg.logistica)

Call:
glm(formula = EsVictima ~ Mes + Dia + Victima + Edad + Genero +
    Nacionalidad, family = "binomial", data = entrenamiento)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2145  -0.0015   0.0000   0.0000   5.1751

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -3.069e+01  8.275e+02  -0.037 0.970414
MesAGOSTO                                1.244e-01  1.002e-01   1.241 0.214438
MesDICIEMBRE                            -2.355e-01  1.061e-01  -2.218 0.026528 *
MesENERO                                -1.436e-01  1.042e-01  -1.379 0.168034
MesFEBRERO                              -9.368e-02  1.037e-01  -0.903 0.366486
MesJULIO                                 1.435e-01  1.001e-01   1.433 0.151855
MesJUNIO                                 2.560e-01  9.943e-02   2.575 0.010030 *
MesMARZO                                -1.003e-01  1.037e-01  -0.967 0.333448
MesMAYO                                  7.287e-02  1.017e-01   0.717 0.473677
MesNOVIEMBRE                            -3.856e-01  1.077e-01  -3.582 0.000342 ***
MesOCTUBRE                              -9.274e-04  1.015e-01  -0.009 0.992706
MesSEPTIEMBRE                            5.091e-02  1.016e-01   0.501 0.616408
DiaJUEVES                               -3.035e-01  8.092e-02  -3.750 0.000177 ***
DiaLUNES                                -1.951e-01  7.914e-02  -2.465 0.013703 *
DiaMARTES                               -2.341e-01  7.920e-02  -2.956 0.003121 **
DiaMI\xc9RCOLES                         -2.586e-01  7.974e-02  -3.242 0.001186 **
DiaS\xc1BADO                             2.608e-01  7.291e-02   3.577 0.000348 ***
DiaVIERNES                               2.223e-02  7.617e-02   0.292 0.770401
VictimaPERSONA                           2.237e+00  6.205e-02  36.050  < 2e-16 ***
VictimaVEHICULO                          1.066e-01  7.559e-02   1.410 0.158629
VictimaVIVIENDA                         -1.012e+00  9.796e-02 -10.329  < 2e-16 ***
EdadDesconocido                          1.584e-02  1.081e-01   0.147 0.883459
EdadMayor de edad                        3.335e+00  8.094e-02  41.203  < 2e-16 ***
EdadMenor de edad                       -4.442e-01  1.199e-01  -3.705 0.000211 ***
GeneroHOMBRE                             4.111e+00  1.335e-01  30.798  < 2e-16 ***
GeneroMUJER                              3.531e+00  1.341e-01  26.333  < 2e-16 ***
NacionalidadALBANIA                      5.174e-02  1.174e+03   0.000 0.999965
```

Figura 21: Resumen de regresión logística. Fuente: Elaboración propia.

En este caso, los “Deviance Residuals” pudieron ser ignorados por tratarse de un modelo de regresión logística; en otros modelos lineales generalizados como regresiones lineales o Poisson podrían ser de relevancia.

Lo importante a rescatar es que se pudo obtener los coeficientes para validar el rendimiento del modelo. Es notable un margen muy grande en los

coeficientes de las distintas variables predictoras; sin embargo, se podrían rescatar algunos atributos que se conectan estadísticamente de manera significativa a la variable dependiente, la cual era si podía ser una víctima de un delito o no. Por ejemplo, los días jueves y sábado tuvieron mayor relevancia sobre los demás días, así como los menores de edad y mayores de edad que tuvieron más relevancia sobre los adultos mayores.

Además, se lograron deducir algunas afirmaciones clave sobre el rendimiento que ha tenido el modelo con el conjunto de datos actual:

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 66097  on 2204269  degrees of freedom
Residual deviance: 16779  on 2204115  degrees of freedom
AIC: 17089

Number of Fisher Scoring iterations: 23
```

Figura 22: Segundo resumen de regresión logística. Fuente: Elaboración propia.

El “Null deviance” y el “Residual deviance” declaran que tan bien se desempeñó el modelo con una constante o con las variables independientes con las que se procesó el conjunto de datos. En este caso, se consiguió evidenciar que existía una reducción estadística en la desviación del modelo al agregar las variables predictoras.

Después de varias iteraciones, logrando ajustar los datos se eliminaron los valores desconocidos y caracteres especiales, y así se logró mejorar un poco los resultados obtenidos:

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54849  on 1094975  degrees of freedom
Residual deviance: 12645  on 1094824  degrees of freedom
AIC: 12949

Number of Fisher Scoring iterations: 23
```

Figura 23: Tercer resumen de regresión logística. Fuente: Elaboración propia.

Los parámetros de desviación se redujeron, así como el “AIC” que funcionó como punto de comparación entre ambas ejecuciones.

## 2. Árbol de decisión

Con base en los parámetros establecidos, se ejecutó el modelo de árboles de decisión con el siguiente resultado:

```
> summary(modelo.arbol)
Call:
rpart(formula = EsVictima ~ ., data = entrenamiento)
  n= 2204270

          CP nsplit rel error    xerror       xstd
1 0.09002169      0 1.0000000 1.0000000 0.01471279
2 0.01000000      4 0.5806941 0.5806941 0.01121655

Variable importance
        Edad      Victima Nacionalidad       Genero
          35           31           20           14
```

Figura 24: Resumen del árbol de decisión. Fuente: Elaboración propia.

Los árboles de decisión mostraron, en esta primera ejecución, un orden de importancia para las variables independientes. En este caso la edad, la víctima, la nacionalidad y el género fueron las más relevantes, respectivamente.

Por otro lado, si se filtraran los datos de tal manera que se pudiera tener mayor completitud de los datos sin utilizar valores desconocidos, los resultados serían muy variables:

```
> summary(modelo.arbol2)
Call:
rpart(formula = EsVictima ~ ., data = entrenamiento)
  n= 1094976

         CP nsplit rel error    xerror       xstd
1 0.1532567      0 1.0000000 1.0000000 0.01544507
2 0.0100000      3 0.5402299 0.5402299 0.01136217

Variable importance
        Edad Nacionalidad      Victima          Dia
          35           32           31            1
```

Figura 25: Segundo resumen del árbol de decisión. Fuente: Elaboración propia.

En este caso, las variables relevantes tomaron otro orden: la edad y la nacionalidad formaron parte importante del algoritmo de clasificación. Esto se vio reflejado en el detalle del modelo:

```
> modelo.arbol2
n= 1094976

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 1094976 4176 0 (0.996186218 0.003813782)
   2) Nacionalidad=AFGANISTAN,ALBANIA,ALEMANIA,ANDORRA,ANTIGUA Y BARBUDA,ARABIA SAUDITA,ARGELIA,ARGENTINA,AR
MENIA,ARUBA,AUSTRALIA,AUSTRIA,BAHAMAS,BANGLADES,BELGICA,BELICE,BERMUDAS,BIELORRUSIA,BOLIVIA,BRASIL,BULGARIA,
CABO VERDE,CAIMAN ISLAS,CAMERUN,CANADA,CHECA REPUBLICA,CHILE,CHINA,CHIPRE,COLOMBIA,CONGO REPUBLICA DEL,CONGO
REPUBLICA DEMOCRATICA DEL,COREA DEL NORTE,COREA DEL SUR,COSTA DE MARFIL,CROACIA,CUBA,CURAZAO,DINAMARCA,DOMIN
ICA,DOMINICANA REPUBLICA,ECUADOR,EL SALVADOR,ESCOCIA,ESLOVAQUIA,ESLOVENIA,ESPANA,ESTADOS UNIDOS,ESTONIA,FILI
PINAS,FINLANDIA,FIYI,FRANCIA,GALES,GEORGIA,GHANA,GRECIA,GUATEMALA,GUAYANA FRANCESA,GUINEA BISAU,GUYANA,HAITI
,HONDURAS,HUNGRIA,INDIA,INGLATERRA,IRAN,IRLANDA,IRLANDA DEL NORTE,ISLANDIA,ISRAEL,ITALIA,JAMAICA,JAPON,JORDA
NIA,KIRGUISTAN,KUWAIT,LETONIA,LIBANO,LIBIA,LIECHTENSTEIN,LITUANIA,LUXEMBURGO,MACEDONIA,MALASIA,MALTA,MARRUEC
OS,MARTINICA,MEXICO,MICRONESIA,MONTENEGRO,NEPAL,NIGERIA,NORUEGA,NUEVA ZELANDA,PAISES BAJOS HOLANDA,PALESTINA
,PANAMA,PARAGUAY,PERU,POLONIA,PORTUGAL,PUERTO RICO,RUMANIA,RUSIA,SALOMON ISLAS,SENEGAL,SERBIA,SIERRA LEONA,S
INGAPUR,SIRIA,SUDAFRICA,SUECIA,SUIZA,SURINAM,TAIWAN,TANZANIA,TRINIDAD Y TOBAGO,TUNEZ,TURQUIA,UCRANIA,URUGUAY
,UZBEKISTAN,VENEZUELA,VIRGENES BRITANICAS ISLAS,VIRGENES DE LOS ESTADOS UNIDOS ISLAS,ZAMBIA 1075342  206 0 (
0.999808433 0.000191567) *
   3) Nacionalidad=COSTA RICA,NICARAGUA 19634 3970 0 (0.797799735 0.202200265)
     6) Edad=Adulto Mayor,Menor de edad 11341  263 0 (0.976809805 0.023190195) *
     7) Edad=Mayor de edad 8293 3707 0 (0.552996503 0.447003497)
      14) Victima=EDIFICACION,VEHICULO,VIVIENDA 4481  841 0 (0.812318679 0.187681321) *
      15) Victima=PERSONA 3812  946 1 (0.248163694 0.751836306) *
```

Figura 26: Modelo de árbol de decisión. Fuente: Elaboración propia.

Lo que se interpretó de este modelo fue que las personas de Costa Rica y Nicaragua, que fueran mayores de edad y que fueran personas físicas, tenían mayor probabilidad de ser víctimas de un delito.

Gráficamente esto se puede representar de la siguiente manera:

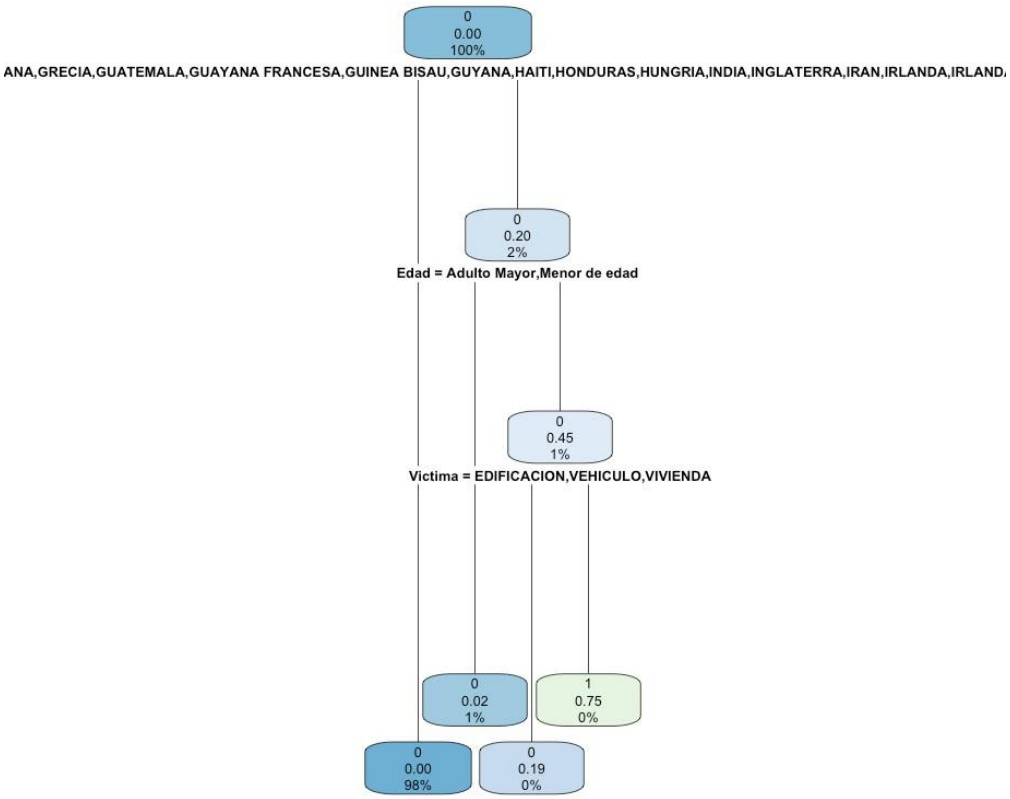


Figura 27: Gráfico del modelo de árbol de decisión. Fuente: Elaboración propia.

### 3. Máquina de soporte vectorial

Los resultados del modelo de máquinas de soporte vectorial se muestran a continuación:

```
> summary(modelo.svm3)

Call:
svm(formula = EsVictima ~ Mes + Dia + Victima + Edad + Genero + Nacionalidad, data = entrenamiento3,
    kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.006578947
```

Figura 28: Resumen del SVM. Fuente: Elaboración propia.

El modelo que redujo el error lo más posible en sus datos de entrenamiento utilizó un costo de 1 y un valor de gamma de 0.5 o menos. Con esto se distinguió como el SVM se comporta prediciendo los valores con las observaciones establecidas.

## 4. Bosque aleatorio

Estos fueron los resultados del modelo de bosque aleatorio:

```
> print(modelo.RF2)

Call:
 randomForest(formula = EsVictima ~ Mes + Dia + Victima + Edad +      Genero + Provincia + Canton + Distrito, data
= entrenamiento,      ntree = 20)
               Type of random forest: classification
                     Number of trees: 20
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0.38%
Confusion matrix:
        0 1 class.error
0 1090700 0           0
1    4175 0           1
```

Figura 29: Modelo bosque aleatorio. Fuente: Elaboración propia.

Cuando el algoritmo seleccionó una muestra con reemplazo para crear un árbol en una iteración, algunas observaciones se quedaron fuera y no se usaron para crear el árbol. Para esas observaciones excluidas del árbol se hizo una

predicción y se calculó el error de la predicción, esto en cada iteración para sacar un error estimado. En este caso fue del 0.38%.

La importancia de las variables según el bosque aleatorio se muestra en el siguiente gráfico:

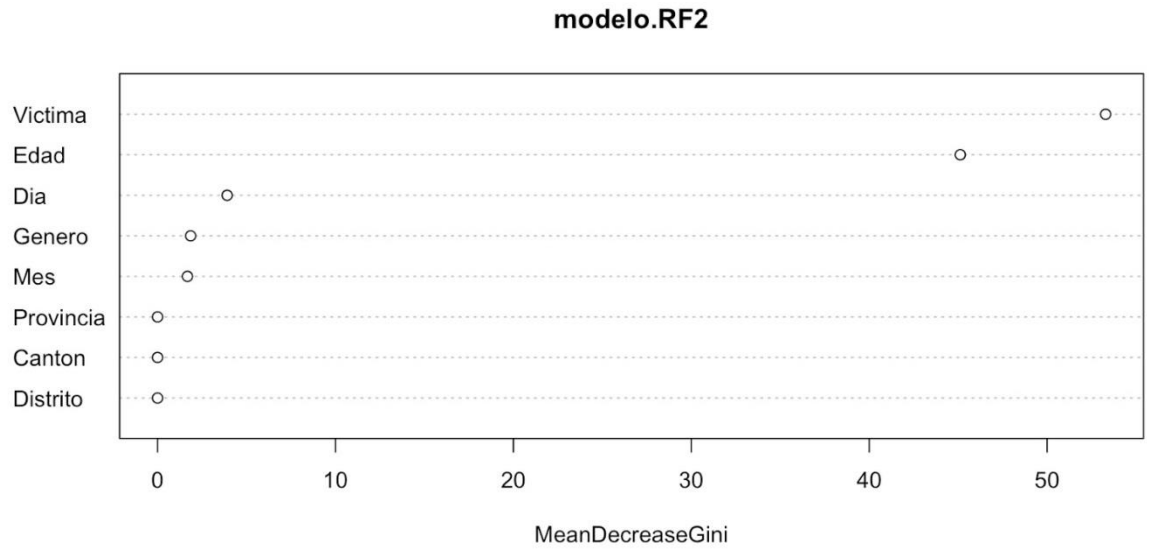


Figura 30: Importancia de variables en el modelo bosque aleatorio. Fuente: Elaboración propia.

## 4.5 Evaluación

En esta fase se evaluó más que la exactitud de los modelos, como por ejemplo si el modelo cumplió con los objetivos del negocio, se determinó si había alguna razón de negocio que hiciera que el modelo no fuera válido y de igual forma se pudo probar al integrarlo con las herramientas del negocio o con casos de la vida real.

### 4.5.1 Evaluación de los resultados

Desde el punto de vista del negocio, se había establecido como criterio de éxito principal el poder estimar la probabilidad de un delito bajo un porcentaje de

fiabilidad aceptable. Con este criterio, a pesar de ser un tanto subjetivo, es inevitable apoyarse principalmente en las métricas de confiabilidad que, desde el punto de vista de minería de datos, los modelos ofrecen, no sólo por ser más específicos y precisos, sino porque también tienen una base objetiva.

Al aplicar las fórmulas mencionadas en la sección 4.4.2 de este capítulo a cada una de las matrices de dispersión de cada modelo se obtuvieron los siguientes valores.

**1. Regresión logística**

```
> table(prueba3$EsVictima,prediccion.log3 >=0.5)

     FALSE   TRUE
  0 467113    373
  1    593   1197
```

Figura 31: Matriz de confusión de regresión logística. Fuente: Elaboración propia.

**Exactitud** = (1197 + 467113) / 469276 = 0.99

**Sensibilidad** =1197 / 1570 = 0.76

**Especificidad** = 467113 / 467706 = 0.99

**Tasa de error** = (593 + 373) / 469276 = 0.00

## 2. Árboles de decisión

```
> table (prueba$EsVictima, prediccion.arbol2[,2]>=0.5)

     FALSE   TRUE
  0 467082    404
  1    569   1221
```

Figura 32: Matriz de confusión de árboles de decisión. Fuente: Elaboración propia.

**Exactitud** = (1221 + 467082) / 469276 = 0.99

**Sensibilidad** = 1221 / 1625 = 0.75

**Especificidad** = 467082 / 467678 = 0.99

**Tasa de error** = (569 + 404) / 469276 = 0.00

Para este modelo se estimó un área bajo la curva de 96%; este índice permitió saber qué tan bueno era el modelo para predecir correctamente una observación seleccionada aleatoriamente.

```
> prediccionROC.ideal <- prediction(prediccion.arbol2[,2], as.logical(as.numeric(prueba$EsVictima)-1))
> as.numeric(performance(prediccionROC.ideal,"auc")@y.values)
[1] 0.9663294
```

Figura 33: ROC de árboles de decisión. Fuente: Elaboración propia.

### 3. Máquinas de soporte vectorial

```
> table(prueba3[,10],prediccion.svm3)
   prediccion.svm3
          0       1
  0 1091497     953
  1    1303    2819
```

Figura 34: Matriz de confusión de SVM. Fuente: Elaboración propia.

**Exactitud** = (2819 + 1091497) / 1096572 = 0.99

**Sensibilidad** = 2819 / 3772 = 0.74

**Especificidad** = 1091497 / 1092800 = 0.99

**Tasa de error** = (1303 + 953) / 1096572 = 0.00

### 4. Bosques aleatorios

```
> table(prueba$EsVictima,predicciones.RF2[,2] >= 0.5)

     FALSE
  0 467486
  1   1790
```

Figura 35: Matriz de confusión de bosques aleatorios. Fuente: Elaboración propia.

**Exactitud** = 467486 / 469276 = 0.99

**Sensibilidad** = 0

**Especificidad** = 467486 / 469276 = 0.99

**Tasa de error** = 1790 / 469276 = 0.00

Los datos desplegados en la siguiente figura muestran de forma resumida los resultados anteriormente calculados:

| Modelo | Regresión logística | Árboles de desición | Máquinas de soporte vectorial | Bosques aleatorios |
|---|---|---|---|---|
| Exactitud | 99% | 99% | 99% | 99% |
| Sensibilidad | 76% | 75% | 74% | 0% |
| Especificidad | 99% | 99% | 99% | 99% |
| Tasa de error | 0 | 0 | 0 | 0 |

Figura 36: Resumen de estadísticas de los modelos. Fuente: Elaboración propia.

### Modelos aprobados

Todos los modelos mostraron un error del 0% y una exactitud cerca del 99%, lo cual es positivo. Para efectos de este proyecto se buscó tener la sensibilidad más alta, ya que es la que muestra el porcentaje de registros positivos que fueron correctamente clasificados; por lo tanto, se tuvo que el modelo de **regresión logística** fue el modelo con **mayor sensibilidad** al tener un **76% de los delitos verdaderos clasificados correctamente**.

Por otro lado, se obtuvo un 99% de especificidad gracias a que se creó una gran cantidad de registros negativos por lo cual los modelos lograron aprender correctamente cuando estos eventos no ocurren o no hay registro de que ocurran.

## 4.6 Reconstrucción del modelo probabilístico

Bajo el esquema analítico, la regresión logística permite estimar la probabilidad de una variable cualitativa binomial en función de una o varias variables predictoras. Una de las principales características de esta clasificación binaria es que se pueden asignar las observaciones en un grupo u otro dependiendo de las variables independientes. Es importante tomar en cuenta que se modela el logaritmo de la probabilidad de pertenecer a cada grupo, la asignación final se hace en función de las probabilidades predichas.

**Predicción de probabilidad**

Dado que el modelo de regresión logística da la flexibilidad de trabajar predicciones lineales en la escala de los predictores agregados, directamente interpretables como *odds* (posibilidades), así como las predicciones de probabilidad de un evento en la escala de la variable de respuesta, se puede obtener la predicción a la etiqueta (0 o 1) dada una combinación de variables. Esto se logró con el atributo type='response', cuando se realizó la predicción:

```
prueba$prediccion.log  <- predict(reg.logistica,newdata = prueba, type = 'response')
```

Estas predicciones en forma de probabilidad se representan en los siguientes gráficos:

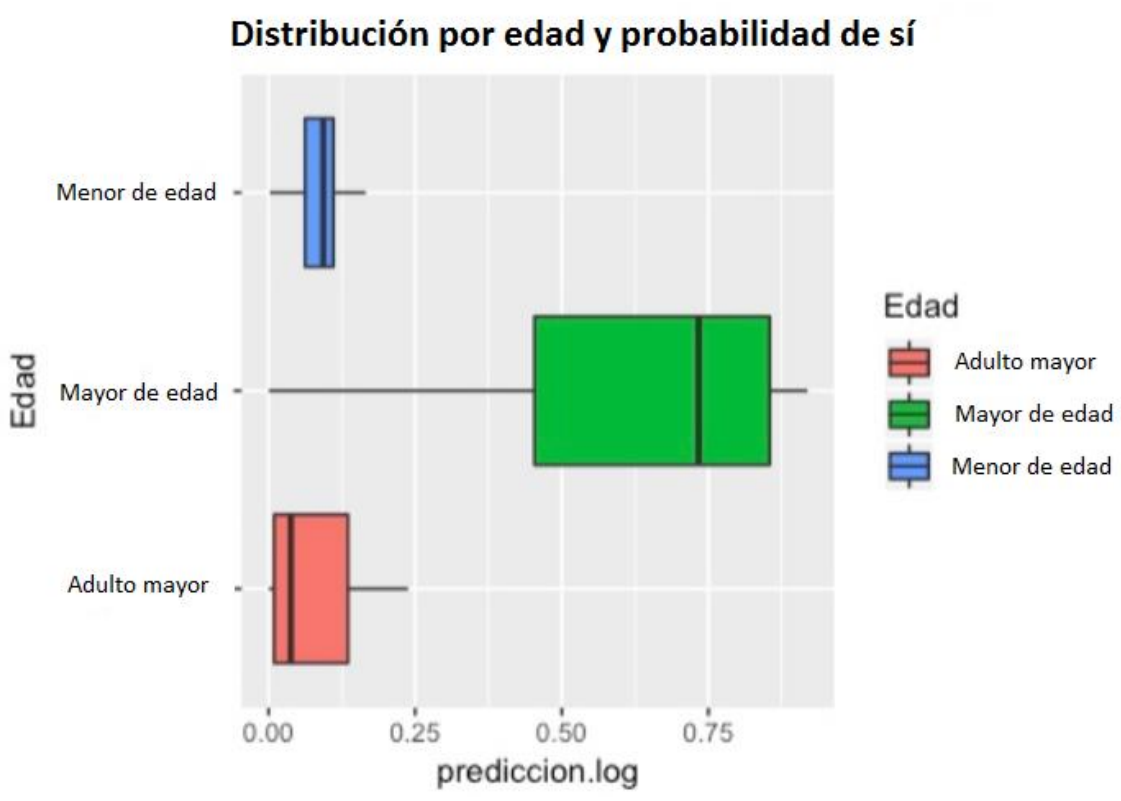


Figura 37: Distribución por edad y probabilidad de sí. Fuente: Elaboración propia.

El modelo hizo las predicciones con base en las variables predictoras y definió una probabilidad para cada combinación de variables, aunque en este caso la mayor parte de las observaciones donde la edad correspondía a mayores de edad, la combinación de los demás factores con los que este atributo estaba combinado dio en su mayoría una probabilidad mayor al 50%.

Sin embargo, esto no define por completo la relación que tiene la edad directamente con el evento, ya que fue un resultado de la predicción y no estrictamente una correlación con la posibilidad de que un evento suceda.

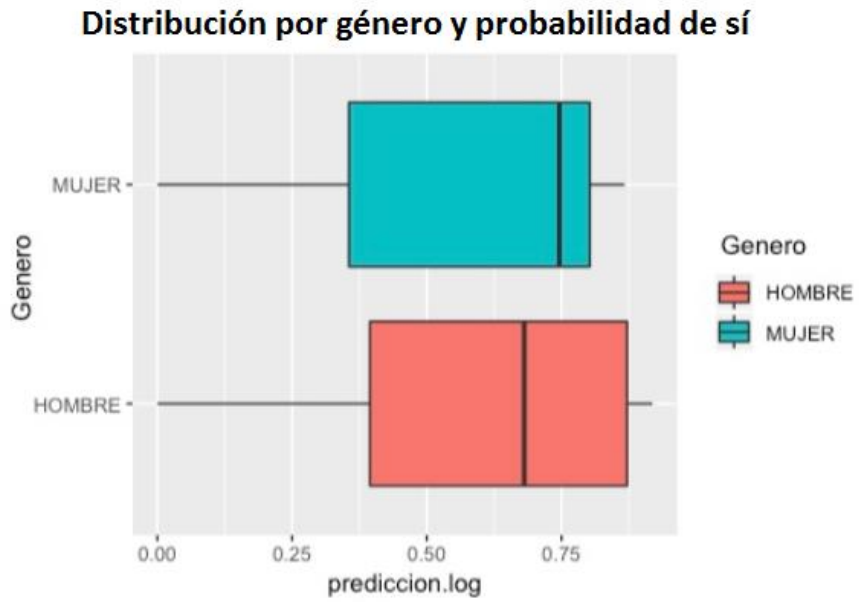


Figura 38: Distribución por género y probabilidad de sí. Fuente: Elaboración propia.

Al agregar más variables a la evaluación se logró ir creando un perfil que pudiera ser representativo, por ejemplo el género, que al igual que a los mayores de edad, las mujeres en promedio fueron clasificadas en la predicción con mayor probabilidad de ser víctimas de un delito, junto con otra combinación de variables de dichas observaciones.

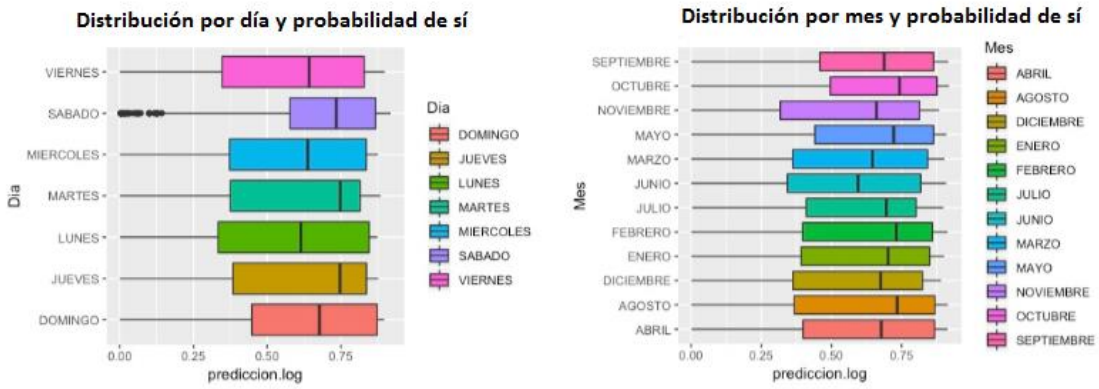


Figura 39: Distribución por tiempo y probabilidad de sí. Fuente: Elaboración propia.

Los días de la semana, así como los meses del año tambien tuvieron ciertas tendencias a raíz de la predicción que se realizó con los parámetros expuestos. Las probabilidades de ser un evento fueron mayores en fines de semana y distribuidas a lo largo del año.

**Predicción de las posibilidades (*odds*)**

La regresión logística asocia la probabilidad de cada valor posible para la variable de respuesta a los predictores. El resultado necesita tener un límite que se acople al resultado de una probabilidad como tal (de 0 a 1); sin embargo, cuando se trabaja con variables de respuesta binaria (es decir, solo dos posibles valores) también se puede calcular la posibilidad (*odds*) de un valor en la respuesta con base en una variable predictora, siendo así:

*Odds > 1, y=1 es más probable*
*Odds < 1, y=0 es más probable*

Como se mencionó en el apartado anterior, se pueden interpretar las predicciones de los valores que hacen más o menos posible que el resultado sea

1 o 0. Esto se logró directamente utilizando el parámetro type='link' o dejando la función *predict* con sus valores por defecto.

```
odds  <-predict(reg.logistica,newdata = prueba)
```

Existen ciertas diferencias, por ejemplo, las predicciones de la probabilidad en el apartado anterior eran estrictamente delimitadas de 0 a 1:

```
> summary(prueba$prediccion.log)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000000 0.0000000 0.0000000 0.0038977 0.0000061 0.9188424
```

Figura 39: Rango de predicción de probabilidad. Fuente: Elaboración propia.

Las posibilidades, en cambio, podían variar entre cualquier valor:

```
> summary(odds)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-30.398 -28.037 -24.836 -20.812 -12.005   2.427
```

Figura 41: Rango de predicción de posibilidad. Fuente: Elaboración propia.

Posterior a esto, se pudo crear un gráfico para interpretar directamente los resultados:

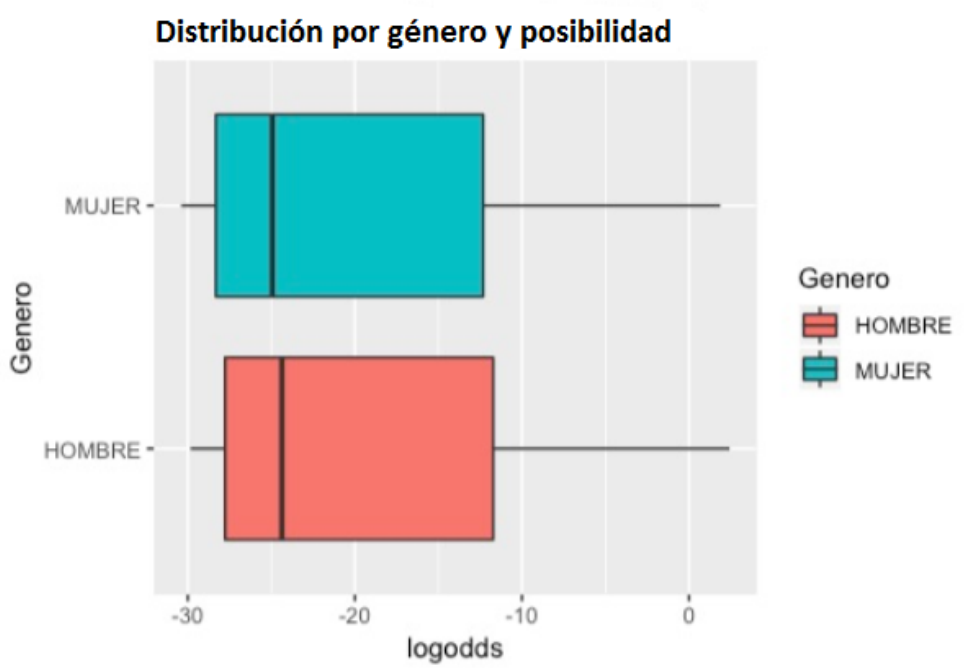


Figura 42: Distribución por género y posibilidad. Fuente: Elaboración propia.

En este gráfico se puede observar que ambas posibilidades eran negativas, lo que indicaría que la posibilidad era muy baja o que influía más en que no fuera posible un delito; no obstante, trabajar directamente con estas posibilidades fue difícil de entender, ya que de primera entrada están en términos logarítmicos. Por ello, fue necesario traducir estas posibilidades logarítmicas a posibilidades reales, utilizando el exponente del logaritmo, y reescribir el gráfico con nuevos datos.

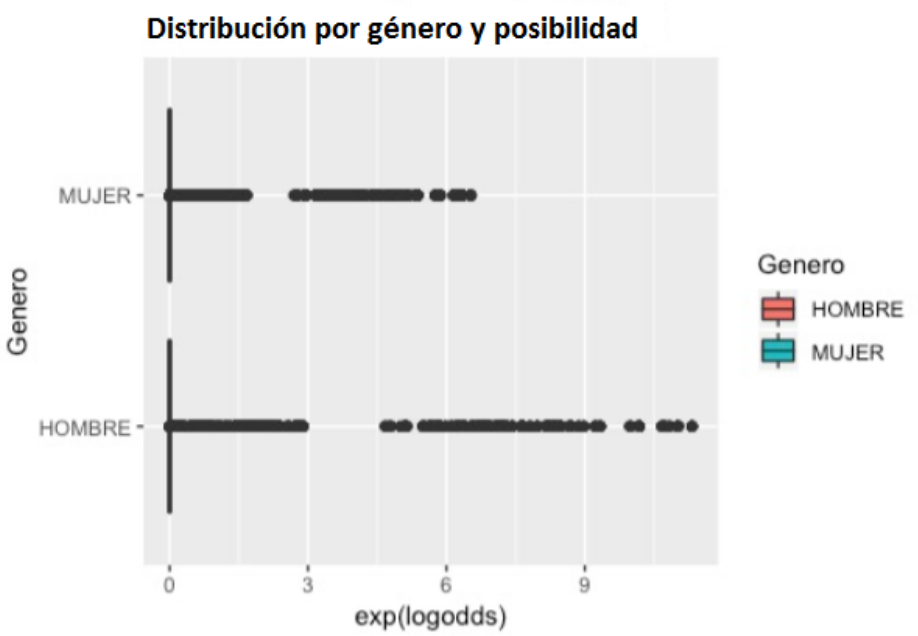


Figura 43: Distribución por género y posibilidad (exponente). Fuente: Elaboración propia.

El gráfico muestra que el género tiene particular influencia en las posibilidades de ser víctima de un delito. El análisis sugirió que cuando el género es ‘hombre’ hay una posibilidad hasta diez veces más alta.

Entre tanto, se muestran otros gráficos asociados al mismo análisis:

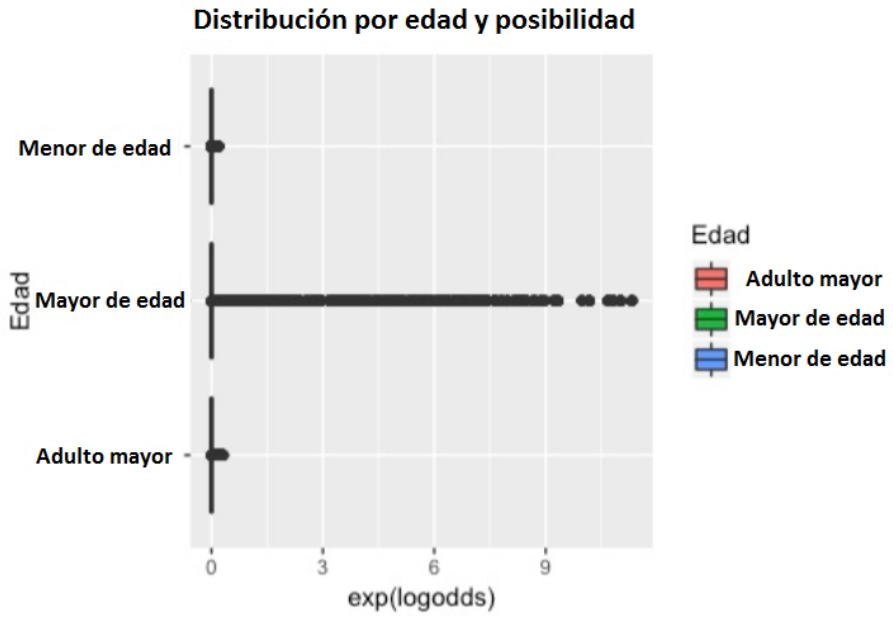


Figura 44: Distribución por edad y posibilidad. Fuente: Elaboración propia.

Nuevamente, las personas mayores de edad tuvieron un valor muy alto en las posibilidades, esto ya representa una asociación directa con la posibilidad de que un delito suceda al pertenecer a este segmento.

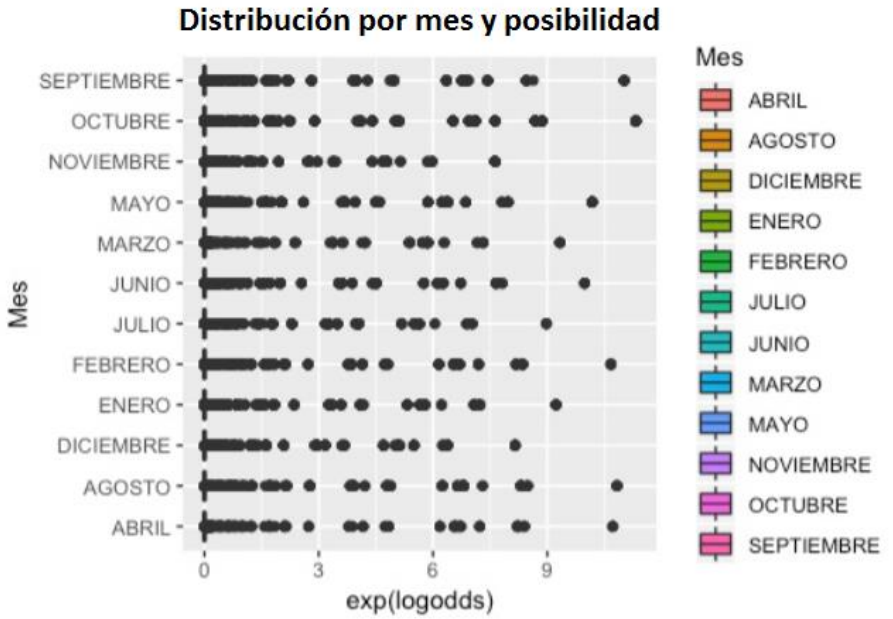


Figura 45: Distribución por mes y posibilidad. Fuente: Elaboración propia.

Relacionado al mes, hay algunas observaciones donde la posibilidad fue relativamente más alta, por ejemplo, en el mes de octubre; sin embargo, el análisis se pudo llevar más a fondo por día de la semana.

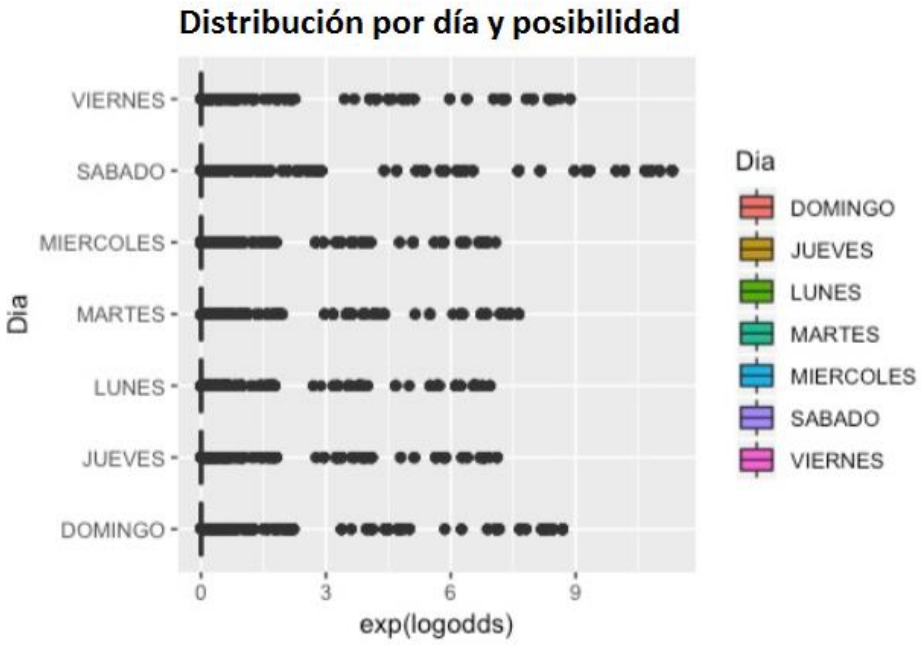


Figura 46: Distribución por día y posibilidad. Fuente: Elaboración propia.

Esto es una representación de las predicciones donde el día de la semana fue influyente para clasificar las observaciones como víctimas de un delito. El sábado se mostró como el día de la semana con mayores posibilidades de que suceda un evento.

**Interpretación de los coeficientes y posibilidades**

El modelo de regresión logística permite también conocer el logaritmo de las posibilidades directamente del resultado del modelo; sin embargo, es recomendable tener un modelo ideal seleccionado de manera progresiva.

El modelo inicial contenía variables que no eran relevantes y que podían generar confusión a la hora de realizar el análisis de los resultados, por ello mejor se seleccionaron las variables relevantes.

```
modelo.ideal1 <- glm(EsVictima ~ Mes + Dia + Victima + Edad + Genero,
                     family= 'binomial',
                     data = entrenamiento)
```

Figura 47: Modelo regresión logística sin nacionalidad. Fuente: Elaboración propia.

La exclusión de la nacionalidad dio cabida a un análisis más específico para así poder precisar variables más concretas que definieran la relación que tenían con la posibilidad de que se diera un evento.

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -8.41895    0.11213 -75.083  < 2e-16 ***
MesAGOSTO          0.02709    0.07495   0.361 0.717795
MesDICIEMBRE      -0.20896    0.07933  -2.634 0.008434 **
MesENERO          -0.08233    0.07693  -1.070 0.284511
MesFEBRERO        -0.01497    0.07576  -0.198 0.843381
MesJULIO          -0.04324    0.07603  -0.569 0.569485
MesJUNIO           0.01219    0.07530   0.162 0.871427
MesMARZO          -0.02859    0.07592  -0.377 0.706509
MesMAYO           -0.13091    0.07778  -1.683 0.092342 .
MesNOVIEMBRE      -0.24067    0.08023  -3.000 0.002701 **
MesOCTUBRE         0.13542    0.07309   1.853 0.063909 .
MesSEPTIEMBRE      0.06382    0.07434   0.859 0.390598
DiaJUEVES         -0.28602    0.06113  -4.679 2.88e-06 ***
DiaLUNES          -0.24745    0.06047  -4.092 4.28e-05 ***
DiaMARTES         -0.20360    0.05983  -3.403 0.000666 ***
DiaMIERCOLES      -0.20280    0.05961  -3.402 0.000669 ***
DiaSABADO          0.29182    0.05303   5.502 3.75e-08 ***
DiaVIERNES         0.02482    0.05631   0.441 0.659429
VictimaPERSONA     2.12957    0.05425  39.254  < 2e-16 ***
VictimaVEHICULO    0.14438    0.06989   2.066 0.038842 *
VictimaVIVIENDA   -1.04743    0.10019 -10.455  < 2e-16 ***
EdadMayor de edad  3.17056    0.07956  39.849  < 2e-16 ***
EdadMenor de edad -0.43971    0.12447  -3.533 0.000411 ***
GeneroMUJER       -0.39304    0.03193 -12.309  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 48: Resultados modelo regresión logística sin nacionalidad. Fuente: Elaboración propia.

Sin embargo, a pesar de los filtros, todavía existían variables que no eran relevantes, el mes por ejemplo tenía muy pocas ocurrencias, por lo que se creó un

modelo en el que solo se contemplaban aquellas variables que sí eran relevantes para obtener mejores resultados.

```
modelo.ideal2 <- glm(EsVictima ~  Dia + Victima + Edad + Genero,
                     family= 'binomial',
                     data = entrenamiento)
```

Figura 49: Modelo regresión logística sin nacionalidad y mes. Fuente: Elaboración propia.

Se obtuvo el siguiente resultado:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -8.45707    0.09988 -84.670  < 2e-16 ***
DiaJUEVES         -0.28467    0.06112  -4.658 3.20e-06 ***
DiaLUNES          -0.24640    0.06046  -4.075 4.60e-05 ***
DiaMARTES         -0.20239    0.05982  -3.383 0.000716 ***
DiaMIERCOLES      -0.20118    0.05960  -3.375 0.000737 ***
DiaSABADO          0.29273    0.05303   5.520 3.38e-08 ***
DiaVIERNES         0.02624    0.05630   0.466 0.641196
VictimaPERSONA     2.12925    0.05425  39.250  < 2e-16 ***
VictimaVEHICULO    0.14437    0.06988   2.066 0.038840 *
VictimaVIVIENDA   -1.04726    0.10019 -10.453  < 2e-16 ***
EdadMayor de edad  3.17029    0.07956  39.846  < 2e-16 ***
EdadMenor de edad -0.43963    0.12447  -3.532 0.000412 ***
GeneroMUJER       -0.39316    0.03193 -12.315  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 50: Resultados modelo regresión logística sin nacionalidad y mes. Fuente: Elaboración propia.

Esta información se socavó a profundidad para poder llevar a cabo un análisis más profundo de las posibilidades.

```
> exp(coef(modelo.ideal2))
     (Intercept)        DiaJUEVES         DiaLUNES        DiaMARTES      DiaMIERCOLES        DiaSABADO        DiaVIERNES
    2.123943e-04     7.522613e-01     7.816127e-01     8.167772e-01     8.177634e-01     1.340083e+00     1.026585e+00
  VictimaPERSONA   VictimaVEHICULO  VictimaVIVIENDA EdadMayor de edad EdadMenor de edad       GeneroMUJER
    8.408553e+00     1.155315e+00     3.508979e-01     2.381433e+01     6.442773e-01     6.749185e-01
```

Figura 51: Interpretación de coeficientes del modelo. Fuente: Elaboración propia.

Las posibilidades de ser víctima de un delito aumentaron en una proporción de 1.34 los días sábado, por otro lado, las posibilidades para mayores de edad eran hasta 23.8 veces más en comparación con otros rangos. Y, contrario a lo que presentaban la mayor parte de las visualizaciones anteriores, la posibilidad aumentó en una proporción de 0.7 para las mujeres en comparación con los hombres.

**Prueba de hipótesis**

En la comparación de género existía una disyuntiva en cuanto a los resultados de las posibilidades de ser víctima de un delito. A pesar de que los hombres aparecían con valores un poco más altos, una prueba de hipótesis ayudaría a rechazar o mantener los criterios preconcebidos que se consideran en una población a partir de una muestra.

Para definir esta prueba de hipótesis se tuvo que definir primeramente la hipótesis nula ($H_0$) como tal:

$$H_0 : m_A \geq m_B$$

Siendo $m_A$ la muestra de los datos para el conjunto de observaciones pertenecientes a los hombres con sus respectivas posibilidades y $m_B$ la muestra de estas para las mujeres. La hipótesis nula exponía que, en promedio, las posibilidades de los hombres eran más altas que las posibilidades de las mujeres.

Con esto se creó la hipótesis alternativa ($H_a$) correspondiente a la contrariedad de la hipótesis nula expuesta anteriormente:

$$H_a : m_A < m_B$$

Calculando la probabilidad y la desviación estándar de la muestra de datos se definieron algunas métricas que ayudaron a describir el error o incertidumbre que se asociaba con las observaciones a la hora de realizar una distribución muestral.

```
# A tibble: 2 x 4
  Genero count  mean    sd
  <fct>  <int> <dbl> <dbl>
1 HOMBRE  1103 0.352  2.20
2 MUJER    687 0.102  1.92
>
```

Figura 52: Probabilidad y desviación estándar de la posibilidad por género. Fuente: Elaboración propia.

La visualización de la muestra de posibilidades también se refería a una diferencia muy superficial:

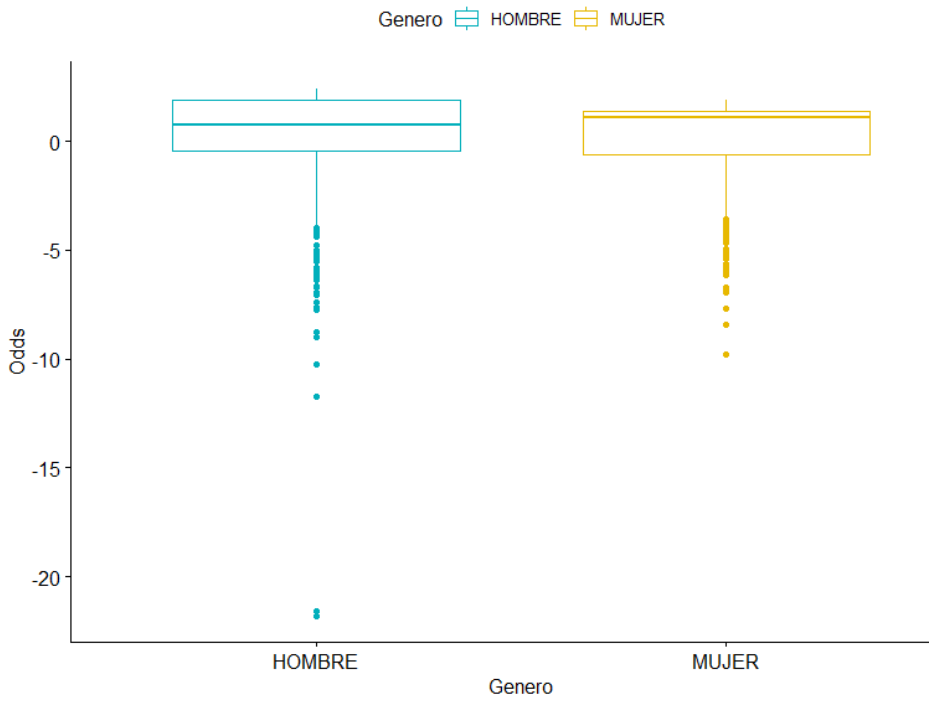


Figura 53: Distribución muestral de la posibilidad por género. Fuente: Elaboración propia.

Preliminarmente existían ciertas suposiciones que se hacían a la hora de correr la prueba de hipótesis, algunas de ellas estrictamente ligadas a la correlación de las observaciones de ambos grupos y al tipo de distribución que mantenían. Las posiciones fueron:

1. Los dos conjuntos de datos son independientes.
2. Los datos de ambos grupos siguen una distribución normal.

3. Ambas populaciones tienen la misma varianza.

Los riesgos de trabajar esta prueba de hipótesis con los datos de toda la población eran que se podía dejar al descubierto alguna variabilidad que no se contempló en los cálculos, por lo que, para ayudar a cumplir los requisitos antes mencionados, se necesitó realizar muestreos aleatorios de manera repetitiva y calcular los promedios de las observaciones para cada género. Con esto la prueba de hipótesis se podía garantizar en términos de independencia de las observaciones, distribución normal y varianza.

Las muestras debían tener la misma distribución de hombres y mujeres del conjunto de datos de la población a estudiar, es decir, que cada muestra en este caso debió seguir una proporción de 62% de hombres y 38% de mujeres.

El tamaño de cada muestra a convenir fue de cien observaciones. Teniendo estos datos se realizó un muestreo repetitivo de cien iteraciones en el cual cada vez que una muestra era extraída del conjunto de datos original, se calculaba el promedio de las posibilidades de cada género, almacenándolas en un conjunto de datos nuevo. Esto dio como resultado dos observaciones por muestra, una de cada género.

Para la primera suposición se consideró de antemano la independencia de ambos grupos. La segunda suposición pudo ser confirmada usando la prueba de normalidad de Shapiro-Wilk donde las hipótesis nulas debían llevar una distribución normal, mientras que las hipótesis alternativas no necesariamente seguían dicha distribución.

```
        Shapiro-Wilk normality test

data:  mean[Genero == "MUJER"]
W = 0.98009, p-value = 0.1351
```

```
        Shapiro-Wilk normality test

data:  mean[Genero == "HOMBRE"]
W = 0.98553, p-value = 0.3465
```

Figura 54: Resultado de la prueba de normalidad de Shapiro-Wilk. Fuente: Elaboración propia.

De los datos de salida, los p-values eran mayores al nivel de significancia de 0.05, lo que implicaba que la distribución de los datos no era significativamente diferente a la distribución normal. En otras palabras, se podía asumir la normalidad.

Con respecto a las varianzas, se realizó un análisis utilizando la librería de F-test de R y se homogeneizaron las varianzas.

```
        F test to compare two variances

data:  mean by Genero
F = 1.1055, num df = 99, denom df = 99, p-value = 0.6187
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7438487 1.6430805
sample estimates:
ratio of variances
          1.105533
```

Figura 55: Resultado de la prueba F-test para la varianza. Fuente: Elaboración propia.

El p-value de F-test es p=0.6187. Era mucho mayor que el nivel alfa de significancia de 0.05. En conclusión, no había una diferencia significativa entre las varianzas de los dos grupos.

Cumplidos los requisitos, se aplicó la siguiente sentencia en R para así computar la prueba, independiente de las suposiciones anteriores:

```
> res <- t.test(mean ~ Genero, data = muestra_final, var.equal = TRUE, alternative = 'greater')
> res

        Two Sample t-test

data:  mean by Genero
t = 6.6352, df = 198, p-value = 1.514e-10
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1819466       Inf
sample estimates:
mean in group HOMBRE  mean in group MUJER
           0.3961970            0.1539046
```

Figura 56: Resultado de la prueba de hipótesis. Fuente: Elaboración propia.

Bajo el contexto de la estadística inferencial, se interpretó el valor de p-value de la prueba de hipótesis por debajo del 0.05, el cual era el criterio de aceptación dado en este proyecto. Se concluyó que las posibilidades de los hombres de ser víctimas de un delito eran mayores al promedio de las mujeres. Cabe destacar que, aunque la hipótesis nula no se rechazó, y dada la naturaleza de la prueba, no se debe asumir que dicha hipótesis sea cierta. Además, si se tuvieran los datos de toda la población (tomando en cuenta que hay crímenes que no se reportaron), no sería necesaria esta prueba de hipótesis, dado que es un tema de muestreo.

### 4.7 Revaluación de los resultados

Más allá de predecir un delito, la extensión de esta investigación buscaba estudiar más a fondo las probabilidades de que sucediera un delito y las variables que hacían que fuera aún más posible; esto con el fin de llevar la exploración de los resultados hacia otras áreas de prevención y focos específicos de atención, dado que en el estrato social sería muy difícil dar con certeza una predicción exacta, al menos con los recursos con los que se contaba para esta investigación.

Estableciendo estos parámetros, se puede recalcar que existen similitudes entre el análisis de las predicciones y las interpretaciones de los coeficientes del modelo de regresión logística.

No cabe duda de que, en el contexto social, las personas mayores de edad, tanto mujeres como hombres, tienen mayor posibilidad de ser víctimas de un delito en comparación con otras edades. En cuanto al día y al mes existe una variabilidad muy trazada en dichos valores, pero los sábados sobresalen y la posibilidad de ser víctima de un delito aumenta ese día.

### 4.7.1 Revisar el proceso

Si bien es cierto que la ejecución del proceso ha resultado según lo previsto, ha habido complicaciones y retos en los cuales se ha tenido que trabajar para lograr los resultados expuestos.

Durante el proceso de manipulación de los datos para su preparación previa al modelado se tuvo que contar con un gestor de base de datos más avanzado para poder modificar los cerca de doscientos mil registros que se contabilizaban en total. Esta labor en un software de hojas de cálculo resultó más complicada de lo que se esperaba.

Por otro lado, la generación de los datos luego de considerar todos los posibles escenarios en todos los distritos disponibles del país daba como resultado un conjunto de datos de más de mil seiscientos millones de registros. Esta cantidad es inmanejable por los programas tradicionales de gestión de base de datos, por lo que se requeriría de software especializado de Big Data.

En cuanto al modelado, se tuvo que esperar varias horas para poder sacar el resultado del modelo de regresión logística (de 3 a 4 horas). Esto puede ser importante de rescatar a la hora de querer hacer predicciones de probabilidad en tiempo real, dada la cantidad de datos que se manejan; la arquitectura de la solución tendría que ser lo suficientemente robusta como para soportar el procesamiento de tal conjunto de datos en un tiempo menor.

Se podría optar por recrear las distintas técnicas de minería utilizando otros modelos de clasificación. Por tanto, con los modelos utilizados se quiso demostrar

las posibilidades que estas técnicas pueden facilitar; sin embargo, el proyecto podría no estar limitado a tales modelos y utilizarse con otros alcances.

### 4.7.2 Determinar los próximos pasos

A partir de aquí, se podrían establecer acciones de crear nuevas iteraciones, ampliar o delimitar el alcance del proyecto, o seguir adelante con la implementación. A raíz de los resultados expuestos en cada uno de los modelos, se pudieron determinar los siguientes pasos para este proyecto. A pesar de que la métrica de sensibilidad del mejor de los modelos no superó el 77%, se considera que puede ser de gran utilidad para abrir las puertas a nuevos proyectos. Por ende, se tomó este resultado como base para iniciar con el proceso de implementación y establecimiento en el ámbito de la seguridad ciudadana.

## 4.8 Implementación

Se planificó el despliegue del modelo, se documentó y se creó una estrategia. Se planeó el monitoreo y el mantenimiento si el modelo va a ser parte de la operación diaria. Finalmente se documentó aquello que surgió bien o mal.

### 4.8.1 Planeamiento de la implementación

Para poder implantar este proyecto sería necesario en primer lugar tener acceso a la base de datos real, es decir, contar con una base de datos que sea actualizada con regularidad. A partir de ahí, los pasos a seguir estarían alineados con los que se especificaron en este documento, desde la comprensión del negocio hasta la implementación.

En segunda instancia, es necesario trabajar de cerca con representantes del Poder Judicial, en cuanto al departamento de gestión de datos abiertos y mejora de procesos, para la utilización de un modelo de minería de datos en los distintos procesos de la entidad para la prevención de delitos y la orientación de esfuerzos de prevención a focos de la comunidad más específicos.

### 4.8.2 Planeamiento de la monitorización y mantenimiento

El monitoreo de la implementación de este proyecto está estrictamente ligado con la revisión periódica de los resultados expuestos por el modelo de minería de datos, dado que, por ser datos originados a partir del comportamiento humano, puede haber cambios en la tendencia y las predicciones pueden arrojar resultados distintos. Es necesario tener un mecanismo de revisión cuatrimestral para evaluar si los resultados del modelo siguen teniendo validez en el ámbito de la seguridad ciudadana.

Es importante tomar en cuenta que dicho modelo necesitará estarse alimentando constantemente de datos nuevos, ya sean hechos ocurridos de delitos o información que le sea útil al modelo para recalcular los resultados con base en los parámetros establecidos.

## Capítulo 5. Propuesta de solución

En esta sección se plantean los distintos escenarios de solución que se consideran adecuados. Uno de los aspectos más importantes a la hora de realizar una propuesta es identificar y comprender el problema que se desea resolver. En este proyecto se especificó en un inicio que la problemática de la seguridad ciudadana estaba creciendo a raíz de factores que, a pesar del esfuerzo del Poder Judicial, se siguen presentando, lo cual afecta directamente en la percepción de la seguridad.

Entre las características de la solución óptima, considerando los resultados expuestos en el modelado de los datos del Poder Judicial utilizando técnicas de minería de datos, está la utilización de la **regresión logística**, la cual se utiliza para predecir la probabilidad de una variable categórica en función de otras variables independientes. Este tipo de modelo tiene mejores resultados según el análisis expuesto en este documento, con una tasa de sensibilidad del 76%.

El modelo de **árboles de decisión** también presentó buenos resultados durante el análisis, y esto corresponde a una idea útil para modelar la probabilidad de un delito en función de otros factores.

Por último, el modelo de **máquinas de soporte vectorial** también se pudo enmarcar en el contexto de clasificación de un evento como delito en correlación con un conjunto de variables predictoras.

# Capítulo 6. Conclusiones y recomendaciones

El objetivo fundamental de este proyecto era la creación de un modelo de minería de datos para la predicción de delitos menores en Costa Rica, que al ser completado pudiera resultar clave para mejorar la seguridad del país.

Inicialmente en este capítulo se presentan las conjeturas a las que se llegaron después de cumplir con los objetivos que se plantearon al inicio del proyecto.

Seguidamente, se exponen las recomendaciones consideradas útiles para perfeccionar el proceso investigativo realizado y que al seguirlas pueden hacer de este proyecto una herramienta útil para mejorar la seguridad ciudadana.

## 6.1 Conclusiones

Con respecto al objetivo específico: “Definir la pregunta que busca ser respondida con el modelo”, es afirmativo que existe una inmensa posibilidad de aplicar modelos de minería de datos a la información histórica que se maneja en el Poder Judicial utilizando técnicas de clasificación. La pregunta que se buscaba responder era si existía la probabilidad de que se llevara a cabo un delito en un momento y lugar determinados contra un perfil específico de víctima.

En relación con el objetivo específico: “Comprender cuáles son los datos necesarios, cómo se recolectarán y su significado preliminar”, se concluyó que entre más detallados sean los datos, mejores van a ser los resultados. La información del lugar y momento específicos, por ejemplo, la fecha y hora de un delito junto con su ubicación en coordenadas podrían ser de ayuda para determinar la probabilidad de un delito de manera más acertada. El perfil de la persona, en un rango de edades o edad específica, podría tener mucho impacto en la definición del tipo de persona que es más vulnerable a estos delitos. Los datos podrían ser recolectados por los distintos canales de denuncia del Poder Judicial y ser almacenados en una base de datos para su posterior análisis.

Sobre el objetivo específico: “Preparar los datos para ser consumidos por el modelo”, se concluyó que, para manejar todos los posibles escenarios a nivel nacional, es necesario contar con recursos informáticos más robustos que sean capaces de manipular grandes volúmenes de datos. De igual forma, fue deducible que los datos requieren algunas transformaciones para poder ser utilizados por los distintos modelos de minería. Esta tarea es de vital importancia para obtener resultados óptimos en los modelos aplicados.

Referente al objetivo específico: “Proponer un algoritmo que se ajuste al problema que se quiere resolver”, se concluyó que existen diferentes técnicas que pueden ser utilizadas para lograr los objetivos que se plantean. En este caso los modelos de clasificación con técnicas de minería de datos con aprendizaje supervisado ayudan a modelar los datos y obtener los resultados acordes a lo que se plantea.

Acerca del objetivo específico: “Evaluar los resultados del modelo”, se concluyó que el modelo de regresión logística aportó resultados más acertados y fiables en cuanto a la probabilidad de que un delito suceda o no. Su importancia recayó en los resultados obtenidos durante la ejecución del modelado donde la sensibilidad del modelo era más alta que la de otros modelos utilizando el mismo conjunto de datos.

En cuanto al objetivo específico: “Desplegar el modelo para que pueda seguir recibiendo datos actualizados y logre cumplir con su función”, se concluyó que se necesita trabajar a fondo con el Poder Judicial para sustentar el modelo con datos más actualizados, implementando la solución a la base de datos de delitos y así poder seguir generando predicciones más acertadas.

Para el Poder Judicial uno de sus principales pilares es la percepción de la población sobre su jurisdicción y la seguridad ciudadana, especialmente en los sectores más vulnerables a la delincuencia y a la criminalidad. El análisis de datos para la estrategia de prevención delictiva y seguridad comunitaria se ha convertido en un importante diferenciador que, consecuentemente, podría aplicarse en áreas

más extensas del Poder Judicial; sin embargo, se necesita empezar a crear iniciativas para el análisis y modelado de estos datos, estudiando a fondo el comportamiento de los mismos para poder llevar a cabo una evaluación más concreta. Asimismo, se requiere el accionar de los analistas de datos en el campo de la minería de datos para poder llevar a cabo estos objetivos. Este proyecto buscaba incentivar nuevos esfuerzos y proveer a la sociedad de información que pueda ser utilizada para su bien.

## 6.2 Recomendaciones

Después de analizar los resultados de la investigación y recabando sobre algunas de las conclusiones expuestas en este documento, se plantean las siguientes recomendaciones cuya finalidad es orientar a quienes busquen la posibilidad de utilizar este trabajo investigativo como base a sus iniciativas de predicción de delitos menores.

Primeramente, se recomienda utilizar un entorno de enfoque al análisis estadístico como lo es R. Esto porque permite la aplicación de modelos predictivos y procesamiento de datos de manera más eficaz. Esta recomendación surge a raíz de que otras distribuciones de software para este propósito requieren el pago de alguna licencia y pueden tener una curva de aprendizaje más compleja.

Respecto a la obtención de datos, se recomienda que sean datos sustraídos del Poder Judicial, ya que existe todo un proceso de recolección y limpieza de datos que ocurre antes de ser publicados, siguiendo los estándares de datos abiertos. Esto se debe a que las calidades de los datos utilizados de otras fuentes tendrían que ser corroboradas y realizarse un estudio previo para validar su veracidad.

Otra recomendación es utilizar la metodología de CRISP-DM, ya que gracias a su naturaleza se muestra exactamente hacia dónde se puede dirigir el proyecto y se encuentra basada en la retroalimentación continua, lo que produce mejores resultados.

Por último, es recomendable utilizar distintos modelos de minería de datos y algoritmos de aprendizaje para poder comparar la confiabilidad entre ellos. En este proyecto se utilizó la regresión logística, árboles de decisión, máquinas de soporte vectorial y bosques aleatorios; no obstante, se puede implementar otro tipo de modelos de clasificación para tener una visión más clara de la tendencia.

## Capítulo 7. Reflexiones finales

Durante muchos años el proceso de minería de datos ha impulsado a los ejecutivos corporativos, gerentes de negocio y a otros usuarios a tomar decisiones de negocios para sus compañías. La utilización de la tecnología para presentar información útil por medio de la recolección de datos, creando informes, cuadros de comparación y visualizaciones siempre había estado orientada a las empresas; sin embargo, los beneficios potenciales de utilizar estas técnicas en el comportamiento humano para mejorar la calidad de vida son cada vez más una realidad.

Los datos históricos, recopilados de sistemas de origen a medida, permiten el análisis de procesos estratégicos a entidades gubernamentales y de bien social para dirigir campañas o esfuerzos de ayuda a las zonas más vulnerables del país. Los modelos de minería de datos combinan un amplio conjunto de técnicas estadísticas, incluyendo la predicción de probabilidades, entre otras, que podrían ser utilizadas en incontables áreas del acontecer diario. Es posible encontrarse con muchos desafíos a nivel de calidad de los datos; sin embargo, se ha encontrado evidencia de generar información valiosa con recursos que ya existen. Este proyecto de investigación lo demuestra.

En Costa Rica existen planes de seguridad, logística policial y capacitaciones en temas de criminalidad, pero la delincuencia sigue golpeando a la comunidad. Se puede pensar en una manera de identificar, mapear y enfrentar estos problemas reduciendo la barrera permisiva del acontecer de estos delitos utilizando metodologías más analíticas y táctico-estratégicas.

El diagnóstico de este trabajo investigativo busca dirigir futuros esfuerzos en materia de predicción de delitos y poder servir como base para iniciativas de seguridad ciudadana.

## Capítulo 8. Trabajos a futuro

Este trabajo se ha propuesto conseguir sus objetivos orientados a la identificación de un modelo de minería de datos para la predicción de delitos menores en Costa Rica y se ha diseñado un esquema investigativo que puede servir de base como contribución para otras líneas de trabajo a futuro.

La primera línea de continuación de este trabajo de investigación es la utilización del Internet de las Cosas (IoT, por sus siglas en inglés) para la sustentación de datos de este modelo en tiempo real. El Internet de las Cosas refiere a la conexión a internet de todo tipo de dispositivos, pasando por un reloj inteligente, por ejemplo; esto con el fin de intercambiar información que posteriormente pueda servir para automatizar en gran medida las posibilidades de uso. Siendo así, sería posible vislumbrar un dispositivo que automáticamente genere una alerta cuando una persona vaya a ser víctima de un delito y así poder tomar las prevenciones del caso. Se puede estimar que el IoT estará en millones de dispositivos en un futuro, lo cual abre las puertas a las posibilidades de trabajar sobre esta misma línea de investigación con un enfoque más aplicativo de dichos modelos de minería en el acontecer diario.

Por otro lado, también se abre la posibilidad de estudio del comportamiento de las redes sociales como entrada a un conjunto de datos nuevo por analizar. Actualmente el trabajo investigativo ha estado orientado a procesar la información de reportes de delito suministrados por el Poder Judicial; sin embargo, las redes sociales dan paso a una posibilidad de conocimiento de delitos en un momento dado. En redes como Twitter, por ejemplo, se pueden observar algunas publicaciones de personas que han sido víctimas de algún delito, esto puede servir como entrada a un modelo de minería que logre analizar estos datos en grandes volúmenes y generar algún tipo de recomendación útil para otros usuarios.

Por último, existe la posibilidad de ampliar esta investigación para tratar temas más delicados como narcotráfico, homicidios o crimen organizado,

buscando siempre el dotar a las entidades pertinentes de herramientas analíticas que puedan ayudar a la mejor toma de decisiones.

## Referencias

Biolchini, J., Mian, P. G., Natali, A. C., & Travassos, G. H. (2005). *Systematic Review in Software Engineering. Systems Engineering and Computer Science Department*.

Carmona, E. J. (2014). *Tutorial sobre Máquinas de Vectores de Soporte (SVM).* Universidad Nacional de Educación a Distancia, Departamento de Inteligencia Artificial, Madrid.

Ernesto A. Rodríguez Moguel. (2005). *Metodología de la Investigación.* Universidad Juárez Autónoma de Tabasco, México.

Roberto Hernández Sampieri, Carlos Fernández Collado, María del Pilar Baptista Lucio. (2010). *Metodología de la Investigación.* México: McGrayHill.

Consejo Nacional de Rectores (Costa Rica). (2017). Programa Estado de la Nación.
*Segundo Informe Estado de la Justicia / PEN* -- 2 ed.

Poder Judicial. (2018). *Datos Abiertos Poder Judicial*. [en línea] Obtenido de: https://www.poder-judicial.go.cr/justiciaabierta/PJCROD_POLICIALES.html [consultado en agosto, 2018].

Dr. Rolando Vega Robert. Corte Suprema de Justicia. (2013). *CONVENIO MARCO DE COLABORACIÓN INTERINSTITUCIONAL ENTRE EL PODER*

*JUDICIAL DE LA REPÚBLICA DE COSTA RICA Y EL MINISTERIO DE SEGURIDAD PÚBLICA.*

[en línea]. Obtenido de: https://www.poder-judicial.go.cr/gica/index.php/noticias/category/49-conveniosnacionales?download=506:convenio-marco-de-cooperacion-interinsticional-entreel-poder-judicial-de-la-republica-de-costa-rica-y-el-ministerio-de-seguridad-publica [consultado en agosto, 2018].

Poder Judicial. (2018). *PRIORIDADES JUDICIALES*. [En línea]. Obtenido de: https://pj.poder-judicial.go.cr/index.php/prensa/193-presidente-de-la-cortesuprema-de-justicia-expone-prioridades-judiciales [consultado en agosto, 2018].

Julio Solís Moreira. *Seguridad ciudadana y prevención de violencia en Costa Rica*. [En línea]. Obtenido de: http://library.fes.de/pdf-files/bueros/fesamcentral/12054.pdf [consultado en agosto, 2018].

Julio Villena Román. CRISP-DM: La metodología para poner orden en los proyectos de Data Science (2016). [En línea]. Obtenido de: https://data.sngular.com/es/art/25/crisp-dm-la-metodologia-para-poner-orden-enlos-proyectos-de-data-science [consultado en agosto, 2018].

CONARE. Programa Estado de la Nación. [En línea]. Obtenido de: https://www.estadonacion.or.cr/estadisticas-index#social [consultado en agosto, 2018].

Poder Judicial. (2018). *Anuarios Policiales*. [En línea]. Obtenido de: https://www.poder-

judicial.go.cr/planificacion/index.php/estadistica/compendioindicadores [consultado en agosto, 2018].

Scimagojr.com. (2018). Data Mining and Knowledge Discovery. [En línea]. Obtenido de: https://www.scimagojr.com/journalsearch.php?q=13579&tip=sid&clean=0 [consultado en junio, 2018].

## Apéndice A

## Código en R de los modelos de datos

```
#Modelo de datos SVM

install.packages('e1071') library(e1071)

datos3 <- read.table("Downloads/Eventos_v3.txt", sep="\t", header=TRUE)
datos3$EsVictima <- factor(datos3$EsVictima)

str(datos3) summary(datos3)

ind3 <- sample(2,nrow(datos3),replace = TRUE, prob = c(0.7,0.3))

prueba3 <-datos3[ind3==1,] entrenamiento3 <- datos3[ind3==2,]

modelo.svm3 <- svm(EsVictima~ Mes + Dia + Victima + Edad + Genero +
Nacionalidad ,data = entrenamiento3,kernel="radial")
prediccion.svm3 <- predict(modelo.svm3, newdata = prueba3[-10], type = 'prob')

str(prediccion.svm3)

table(prueba3[,10],prediccion.svm3)

#Modelo de datos regresión logística

install.packages('caTools')

library (caTools)
```

```
 getwd()

set.seed(1234) splt3 <- sample.split(datos3$EsVictima, SplitRatio = 0.7)
entrenamiento3 <- datos3[splt3,] prueba3 <- datos3[!splt3,]

str(entrenamiento3) summary(entrenamiento3) str(prueba3) summary(prueba3)

reg.logistica3 <- glm(EsVictima ~ Dia + Victima + Edad + Genero+Nacionalidad,
family= 'binomial',                data = entrenamiento3)
                                                                    exp(coefficients

prediccion.log3  <- predict(reg.logistica3,newdata =prueba3, type = 'response')

str(reg.logistica3) summary(reg.logistica3) summary(prediccion.log3)

table(prueba3$EsVictima,prediccion.log3 >=0.5)

library(ROCR) prediccionROC.idel <-
prediction(prediccion.log3,as.logical(as.numeric(prueba3$EsVictima)-1))
as.numeric(performance(prediccionROC.idel,"auc")@y.values)

#Modelo de datos bosque aleatorio install.packages('randomForest')
library(randomForest)

modelo.RF2<-randomForest(EsVictima ~
                Mes + Dia + Victima + Edad + Genero + Provincia + Canton +
Distrito ,          data=entrenamiento3,          ntree=20)
```

```
predicciones.RF2<-predict(modelo.RF2,newdata = prueba3,type = 'prob')


getTree(modelo.RF,1) head(predicciones.RF)
table(prueba$EsVictima,predicciones.RF[,2] >= 0.5)
table(prueba$EsVictima,predicciones.RF2[,2] >= 0.5)


cforest(EsVictima ~ ., data=modelo.RF) getTree(modelo.RF, 1, labelVar=TRUE)


#Modelo de datos árboles de decisión install.packages('rpart.plot')
install.packages('fancyRpartPlot') install.packages('RColorBrewer') library (rpart)
library (rpart.plot) library(RColorBrewer) install.packages('rattle') library(rattle)


modelo.arbol2 <- rpart(EsVictima ~ . , data = entrenamiento3 prediccion.arbol2 <-
predict(modelo.arbol2, newdata = prueba3 type = 'prob')


summary(modelo.arbol2) prp(modelo.arbol2)


head(prediccion.arbol) table (prueba$EsVictima, prediccion.arbol2[,2]>=0.5)


fancyRpartPlot(modelo.arbol2)  rpart.plot(modelo.arbol2)



prediccionROC.ideal <- prediction(prediccion.arbol2[,2],
as.logical(as.numeric(prueba3$EsVictima)-1))
as.numeric(performance(prediccionROC.ideal,"auc")@y.values)
```