Modelo predictivo aplicado a las Estadísticas Policiales del Poder Judicial de Costa Rica para el análisis de la criminalidad por regiones

Dennis Coto Leiva,
Escuela de Ingeniería,
ULACIT,
San José, Costa Rica
dcotol848@ulacit.ed.cr

Franklin Montero Valverde,
Escuela de Ingeniería,
ULACIT,
San José, Costa Rica
fmonterov916@ulacit.ed.cr

Natalia Avilés Fioravanti,
Escuela de Ingeniería,
ULACIT,
San José, Costa Rica
navilesf387@ulacit.ed.cr

Julio Córdoba Retana,
Escuela de Ingeniería,
ULACIT,
San José, Costa Rica
jcordobar022@ulacit.ed.cr

Resumen – Por medio del análisis descriptivo realizado a los Datos Abiertos del Poder Judicial de Costa Rica, se reflejan números alarmantes en cuanto a la cantidad de denuncias impuestas en el Organismo de Investigación Judicial, superando las cincuenta mil denuncias en el año 2019. Con estos números en mente, el objetivo de esta investigación es generar un modelo de análisis de datos que permita potencializar estas estadísticas y señalar anticipadamente las regiones con mayor propensión de sufrir delitos en el próximo quinquenio, con el fin de promover la proactividad, tanto del ciudadano como de la policía para estar alerta y evitar los próximos delitos. Se aprovechan los modelos de predicción estadística mediante métodos matemáticos comprobados y aplicables a los datos obtenidos y su comportamiento durante los años 2015-2019. El análisis realizado refleja la necesidad de aplicar el algoritmo de regresión lineal simple el cual es aplicado a la solución desarrollada que está disponible para todos los costarricenses en el sitio Tableau Public. Los resultados muestran predicciones negativas para el país, especialmente en el Gran Área Metropolitana, el comportamiento de los delitos impactará significativamente esta zona, lo cual indica la necesidad de establecer programas de fortalecimiento policial, mejoras en la educación y el empleo para contrarrestar los potenciales delitos proyectados.

Palabras clave – predicción de delitos, patrullaje preventivo, estadísticas policiales, lucha contra el crimen.

Abstract– Through the descriptive analysis carried out on the Open Data of the Costa Rican Judicial Power, alarming results are reflected in terms of the number of complaints imposed in the Judicial Investigation Organism (OIJ), exceeding fifty thousand complaints in 2019. Based on those numbers, the objective for this research is to generate a data analysis model that allows to potentiate these statistics and to indicate in advance the regions with the greatest propensity to suffer crimes in the next five years, to promote the proactivity of both the citizen and the police to be alerted and to avoid upcoming crimes. Statistical prediction models are used using proven mathematical methods applicable to the data obtained and their behavior during the years 2015-2019. The analysis carried out reflects the need to apply the simple linear regression algorithm which is applied to the developed solution that is available to all Costa Ricans on the Tableau Public website. The results show negative predictions for the country, especially in the Greater Metropolitan Area (GAM), the behavior of crimes will significantly impact this area, which indicates the need to establish police strengthening programs, improvements in education and employment to counter the potential crimes projected for the next five years.

Keywords -- crime prediction, preventive patrol, police statistics, crime fighting.

I. Introducción

En la década de 2010, la percepción ciudadana sobre la seguridad era un tema sonado en la agenda política y en el marco electoral costarricense [1]. Sin embargo, hoy, la ciudadanía plantea otras cuestiones de carácter más urgente. Esta investigación señala si se ha tenido mejora, o no, en materia de seguridad en el último quinquenio, usando como fuente de datos las estadísticas policiales publicadas de manera abierta en el sitio web del Poder Judicial costarricense. Además, se plantea una solución tecnológica que fortalezca a la sociedad y transforme las falencias identificadas (2021).

Según el Estado de la Nación [2], de acuerdo con las denuncias netas recibidas en el sistema de justicia entre los años 2015 y 2019, se muestra una cantidad significativa, por ejemplo, en el año 2019 se superan las cincuenta y siete mil denuncias anuales, y en el 2016 más de cincuenta y ocho mil. Aunado a esto, la misma fuente demuestra la tasa de delitos dolosos contra la vida por cada cien mil habitantes entre los años 2000 y el 2009, con un promedio de 223 casos anuales, mientras que entre los años 2010 y 2019 la misma tasa refleja un promedio de 249 delitos por año, lo que muestra un crecimiento de aproximadamente un 11.7% en la década de 2010.

Uno de los principales factores para evaluar este tema, es el trabajo realizado por organizaciones públicas para cumplir con su deber moral de rendición de cuentas por medio de la apertura de datos e información, y el deber cívico de fiscalizar a las instituciones gubernamentales en su accionar cotidiano, un ejemplo de cómo el uso de datos abiertos le proporciona a la ciudadanía herramientas para calificar el trabajo de dichas instancias.

El conjunto de datos seleccionado permite realizar un análisis estadístico geográfico, por género, por nacionalidad e histórico, entre otros; lo cual posibilita correlacionar diversos factores cualitativos para acompañar el análisis cuantitativo, y así ilustrar las diferencias de la segmentación de datos. Sin embargo, al triangular otras fuentes y aplicar modelos estadísticos de predicción y análisis de datos, se señalan las

posibles tendencias de la criminalidad en las distintas regiones del país.

Actualmente, la tecnología se ha aliado a modelos estadísticos para el análisis de datos, y de esta manera, automatizar tareas que tomarían mucho tiempo y recursos. En este trabajo se identifican modelos de análisis correlacional y herramientas tecnológicas que permiten crear una solución que apoye la toma de decisiones para el accionar proactivo tanto policial como ciudadano, mediante proyecciones predictivas en materia de criminalidad en Costa Rica.

Para ello, se utilizan los datos abiertos del Poder Judicial costarricense para pronosticar potenciales delitos en las zonas con mayor criminalidad del país a lo largo de un quinquenio. La primera hipótesis por analizar en esta investigación es si la tasa de criminalidad en Costa Rica entre los años 2015-2019, es mayor en cantones urbanos que en rurales, concentrándose una mayor cantidad de delitos en el Gran Área Metropolitana. Otras dos hipótesis por comprobar son si la criminalidad en Costa Rica se ha mantenido o incrementado en el último quinquenio, a pesar de que la percepción de seguridad ciudadana se mantiene sin variantes. Por otro lado, la tercera hipótesis es que la seguridad ciudadana se verá afectada en el próximo quinquenio, con un crecimiento sostenido de los delitos.

Este articulo tiene el fin de contestar las hipótesis mencionadas anteriormente, mediante la extracción y análisis de datos relevantes que permitan apoyar a la sociedad civil para que esta colabore con las instituciones encargadas de la seguridad ciudadana. Se plantea un modelo de análisis de datos que permita potencializar las estadísticas policiales del sitio web de Datos Abiertos del Poder Judicial de Costa Rica [3], y señalar de manera anticipada las zonas con mayor propensión a sufrir delitos.

Según lo mencionado, los datos de estadísticas policiales permiten obtener información histórica sobre los delitos y subdelitos en Costa Rica, por género, cantón y tipo de víctima.

El análisis de la información se realiza en el periodo 2015-2019, basado en las últimas estadísticas publicadas en la plataforma. La inclusión del cantón de Río Cuarto se toma a partir de su fundación en abril del 2018 [4], previo a esto no existen datos de la zona.

II. Metodología

Este artículo aborda la investigación desde el enfoque cuantitativo, generando un modelo de análisis de datos que permite triangular las métricas presentes en el sitio de datos abiertos de Poder Judicial [3] y los datos de regionalización del MIDEPLAN, con el objetivo de predecir la propensión de criminalidad en una zona específica del país, en aras de contribuir a que la ciudadanía visualice y comprenda la evolución de los delitos en las regiones del país.

La información utilizada para el análisis y desarrollo del artículo se encuentra disponible para todos los ciudadanos en el sitio web del Poder Judicial [3]. Actualmente, en este sitio se pueden encontrar datos públicos en diversos formatos, como Excel, JSON y txt. La información pertinente a la regionalización, cantones, distritos y cantidad de habitantes por kilómetro cuadrado es documentada por el Ministerio de Planificación Nacional (MIDEPLAN). Estos datos, a diferencia de los ofrecidos por el Poder Judicial, se encuentra en mapas en formato PDF y documentos de Excel como tablas finalizadas, por lo que se requiere un proceso de recolección y limpieza de estos, con involucramiento de la parte manual o procesos automatizados más complejos.

Se selecciona un diseño indagatorio no experimental longitudinal de tendencia, utilizando las herramientas de Tableau 2020.4 para la visualización de datos, Tableau Public para brindarle los resultados al público y el lenguaje de programación Python en la versión 3.9.7, para el análisis predictivo por medio del algoritmo de regresión lineal.

El análisis sigue las etapas mencionadas por Liebowitz [5], comenzando por la recopilación de datos; posterior a su obtención y limpieza, se procedió a utilizarlos en la herramienta Tableau 2020.4, para unir la información y mostrar el análisis de datos de una manera visual y dinámica, siendo estos publicados en el sitio de Tableau Public como herramienta para el consumo abierto del público. El análisis predictivo se realiza mediante Excel y la programación de una herramienta con el lenguaje Python; estas arrojan resultados que permiten responder a las interrogantes de predicción planteadas en el artículo. Para generar estas predicciones, se analiza el comportamiento de los datos con el fin de plantear el algoritmo de predicción más apto y de esta manera obtener resultados con un nivel de confianza mayor al 65%. Se concluye que el algoritmo de regresión lineal es el que mejor se adapta, de acuerdo con el comportamiento de los datos de manera continua en el tiempo con respecto a la cantidad de delitos perpetrados.

III. Marco Teórico

Esta investigación utiliza datos abiertos como base de su análisis, entiéndase por datos abiertos aquellos “que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen.” [6]

Los datos abiertos se han convertido en una herramienta que le permite a la ciudadanía fiscalizar el actuar del gobierno. En el año 2015, un grupo de gobiernos alrededor del mundo, junto con organizaciones de la ciudadanía, acuñan este concepto en el “Open Data Charter” [7] (Carta de los Datos Abiertos), donde se definen seis principios que corresponden con esta filosofía:

Son abiertos por defecto, son oportunos y completos, son accesibles y usables, son comparables e interoperables, mejoran la gobernanza y la participación ciudadana y, finalmente, se utilizan para la innovación y el desarrollo inclusivos.

Costa Rica es parte de las 85 naciones que adoptan estos principios desde octubre del 2016 [8]; sin embargo, según el informe más reciente de la Comisión Económica para América Latina y el Caribe (CEPAL) [9], el país es uno de los pocos en el istmo sin legislación de acceso a la información pública,

poniéndonos en el mismo grupo de Cuba, Venezuela, Bolivia y Haití. Gracias a la constante acción de organizaciones de la sociedad civil, se tiene el Decreto Ejecutivo 40199, publicado en la Gaceta, San José, Costa Rica, viernes 12 de mayo del 2017, donde se indica que el gobierno central tiene el deber de implementar la sección de acceso a la información en sus plataformas digitales.

Desde ese entonces, instituciones como el Poder Judicial de Costa Rica han hecho un intento por brindarle a la ciudadanía información transparente y fácil de utilizar, creando un portal de Datos Abiertos que ofrece múltiples estadísticas que van desde presupuestos institucionales y salarios, hasta información relacionada con violencia doméstica, feminicidios y otras estadísticas relevantes. De esta manera, la institución traslada a la sociedad civil la responsabilidad de utilizar y consumir esta información para fiscalizar y coadyuvar al gobierno con el análisis de datos institucionales, y brindar herramientas que posibiliten la solución de problemas que enfrenta la sociedad costarricense.

Entre las opciones que brinda el portal de Datos Abiertos de esta entidad, destacan cantidad de estadísticas policiales donde se brinda la información de los delitos cometidos a lo largo y ancho del país desde el año 2015, incluyendo delitos como robos y tachas a vehículos; asaltos, homicidios, entre otros. Estas mismas estadísticas señalan la fecha en que se cometió el crimen, el género y la nacionalidad de su perpetrador(a) y quién fue la víctima, todo de manera confidencial, sin incluir nombres, cédulas, placas ni ninguna otra información que permita deducir la identidad de las partes involucradas.

Otros datos pertinentes incluyen la división por provincia y cantón, de manera que se determina el punto geográfico donde se perpetró el delito. Como toda fuente de datos, tiene sus limitaciones; por ejemplo, en este caso se omiten las horas en que se cometió el delito, así como la edad de las partes; además, el Manual de Usuario de las Estadísticas Policiales [10] indica que: "Los datos expuestos en la consulta toman como fuente las denuncias interpuestas directamente ante el OIJ; además, no incluyen denuncias de Fiscalías u otras Policías; y, finalmente, la fecha que se toma como referencia para el cómputo del delito (salvo contadas excepciones), es la fecha del hecho y no la fecha de la denuncia."

También es importante definir tres conceptos clave que utiliza esta fuente de datos:

El primer concepto es delito; entendido como una categoría de las estadísticas policiales utilizada para señalar en el nivel macro, el tipo de infracción penal perpetrada. Se dividen en: asalto, hurto, robo, tacha de vehículo, robo de vehículo y homicidio. Según el Manual de Usuario de las Estadísticas Policiales "una categoría delictiva policial (por ejemplo, asalto) puede convertirse tiempo después (horas, días, semanas o meses inclusive) en un homicidio, debido al fallecimiento de la persona, producto de las heridas infligidas." [10].

El segundo es el subdelito, el cual se considera una subcategoría de los delitos, donde se puede expandir en detalle una descripción predeterminada del delito cometido. Por ejemplo, para asaltos se tiene entre subdelitos: arma blanca, arma de fuego, arrebato, golpes, inmovilización, entre otros. Todos están acompañados de características comunes que permiten agrupar el delito denunciado.

Finalmente, el tercer y último concepto es víctima; el cual se indica que "puede ser una persona física o jurídica, entre otros. No obstante, policialmente interesa conocer hacia quién o hacia qué dirige su actividad el delincuente, por ejemplo, para la consulta la víctima puede ser tanto una persona como un bien material, ya sea mueble o inmueble." [10]

El siguiente paso es utilizar una técnica de regionalización que permita ilustrar una estrategia nacional, como suele ser utilizado por estudios estadísticos de entidades como el Instituto Nacional de Estadística y Censos (INEC). El ente encargado de definir esta regionalización es el MIDEPLAN, elaborado con el objetivo de agrupar distritos que cumplen con características topográficas y socioeconómicas similares, lo cual permite enfocar planes de acción acordes con cada zona. Esta división será uno de los pilares del estudio, a continuación, se detalla la lista de las seis regiones por analizar: Central, Brunca, Chorotega, Huetar Norte, Pacífico Central y finalmente la región Huetar Caribe.

Este tema no es ajeno al trabajo de múltiples instituciones gubernamentales, y aunque a veces pase inadvertido, según lo demuestra el último informe de la Encuesta de Percepción de Seguridad en Costa Rica (IDESPO, 2019) [11], un "68,9% de la población considera no vivir de manera segura en el país, y un 31,1% considera que sí. Es decir, según esta fuente, la cantidad de personas que se sienten inseguras duplica a la de las que se sienten seguras.", tal como apuntan estudios como el realizado por la Escuela de Estadística de la Universidad de Costa Rica (UCR) en el 2012 [1], donde un 59% de los entrevistados percibía un aumento de la criminalidad en nuestra nación.

En cuanto a temas demográficos de ubicación urbana o rural, la diferencia es apenas de un punto porcentual entre ambos, siendo la percepción de inseguridad mayor en la zona urbana, pero sin diferencias significativas que motiven a enfocar el estudio en una comparativa de criminalidad urbano-rural. Por su parte, la (IDESPO, 2019) detalla ciertas diferencias de percepción de seguridad según el género, donde las mujeres manifiestan mayor sensación de inseguridad, como suele ser normal en este tipo de estudios debido a factores como el acoso callejero.

La encuesta de actualidades, realizada en un ámbito preelectoral busca entender, además de la seguridad, las pasiones que movían a la población antes de las elecciones del 2014, ya que en 2010 el tema de seguridad había sido un punto importante de discusión en los debates políticos. El IDESPO indica que "pese a la disminución reciente de la inseguridad en el país, la percepción de la ciudadanía es que durante los últimos tres años continúa aumentando" [1], por lo que del análisis se puede inferir que independientemente de la labor realizada por el Ministerio Público, el Poder Judicial, el Ministerio de Justicia

y Paz en cuanto a programas de prevención y políticas de ejecución, la percepción de inseguridad obedece a otros factores como: medios de comunicación, experiencias personales y de conocidos, y ámbito de esparcimiento donde se desenvuelva; tal como lo menciona el informe del IDESPO "la percepción es un proceso de construcción individual pero también influyen elementos de naturaleza grupal." [11]

Este estudio utiliza la información brindada por el Poder Judicial, tomando como base las denuncias de los delitos perpetrados en el último quinquenio, y así tener una óptica objetiva la próxima vez que en medios de comunicación nacionales e internacionales, se haga referencia a conceptos como la percepción de inseguridad frente a la cantidad de delitos cometidos.

Para introducir el contexto sobre el cual se basa la investigación, se establece una aproximación teórica a los conceptos que son parte del análisis. Trujillo [12] destaca las drogas como uno de los principales motivadores para llevar a cabo los crímenes, y la posesión de armas como herramienta que ha ganado popularidad en las últimas décadas por los delincuentes.

Como expone Alvarado [13], a partir de los años ochenta Costa Rica da sus primeros pasos para fortalecer la sociedad mediante diferentes estrategias que prevengan el delito. Durante el gobierno de Oscar Arias Sánchez (1986-1990), se establece un Plan Nacional de Desarrollo, donde se destacan dos objetivos. El primero orientado a fortalecer y consolidar los programas preventivos en materia de seguridad, el segundo se enfoca en la formulación de proyectos de prevención integral, impactando especialmente a la juventud presente en las áreas donde se identifiquen ciertas conductas patológicas. A partir de esta iniciativa, los siguientes gobiernos proponen mejoras a este Plan Nacional, siempre tomando en cuenta factores que favorezcan las acciones proactivas en contra del crimen.

Por otro lado, de acuerdo con Villalobos [14], se cuenta con algunas referencias relativas al empleo de herramientas tecnológicas para el abordaje de situaciones como la expuesta; tal es el caso del Banco Interamericano de Desarrollo (BID). Dicha institución ha puesto a prueba Sistemas de Información Geográfica, Sistemas de Big Data y Sistemas de Monitoreo de Circuito Cerrado (CCTV) como apoyo a las estrategias policiales orientadas a la planificación, prevención del delito e investigación y persecución de presuntos delincuentes.

Las tecnologías emergentes están marcando el camino hacia la cuarta revolución industrial. Como indica Arteaga [15], a pesar de su implementación en los países, organizaciones o personas, esta requiere de esfuerzo, tiempo y recursos; también ofrece una gran oportunidad de crecimiento y mejoras debido al impacto innovador que genera en las estrategias y procesos.

Entre las tecnologías que apoyan esta investigación están la analítica de datos, el aprendizaje automático o machine learning y la computación en la nube. Como menciona Liebowitz [5], tanto la analítica de datos como la inteligencia artificial trabajan de la mano hacia al beneficio de la ciencia, la salud, la economía, la manufactura y por supuesto, seguridad ciudadana.

Con respecto a la analítica de datos como pilar fundamental hacia la generación de predicciones, Rábade [16] expone un concepto implementado en ciudades como Los Ángeles, California, conocido como "patrullaje de predicción". Este se basa en la recopilación y análisis masivo de datos, donde se genera información crucial para la policía, indicando qué delitos son probables, dónde y cuándo, en una zona de aproximadamente 45 metros cuadrados; de tal manera que se logre intensificar la vigilancia policial en estas zonas. Como resultado de estas acciones, la criminalidad se ha reducido en un 13%, lo cual es significativo en un territorio de más de 1.3 millones de personas.

Hay dos conceptos importantes que también destaca Rábade [16], uno de estos es la "actuación policial orientada a los problemas" y el otro es la "actuación policial basada en materia de inteligencia". El primer concepto se orienta hacia las estrategias donde la policía ataca aquellos problemas recurrentes mediante acciones como el trabajo policial comunitario, la cooperación con otras instituciones como escuelas, organizaciones cívicas y centros comunitarios. El segundo concepto destaca la importancia de la información para generar conocimiento que apoye las decisiones de las autoridades. Cabe destacar que tanto la actuación policial orientada a problemas como la actuación basada en inteligencia no son excluyentes; esta última apoya actividades comunitarias y de vigilancia, para así dirigirse con mayor propiedad y fundamentos hacia las zonas más problemáticas.

En el campo de la analítica de datos es importante explicar otros conceptos. La ciencia de datos se compone de ciertas tareas para la manipulación de estos, lo que lleva a un resultado final, con datos preparados para un posterior análisis. Según Liebowitz [5], los pasos que forman parte de este concepto son: generación, recolección, procesamiento, gestión, recuperación y análisis. El propósito de estas acciones es generar datos limpios que hagan más sólido el análisis y generación de resultados. Los datos limpios pasan por un proceso donde se identifican los faltantes, repetidos o que tienen algún conflicto con otros datos disponibles.

Cuando el proceso de limpieza ya se ha realizado, aparece el perfil de científico de datos. Este colaborador extrae todo aquello que realmente es valioso para el análisis, además, identifica elementos no disponibles que podrían afectarlo. Estas actividades no son fáciles; afortunadamente, la inteligencia artificial es una herramienta que ha avanzado en gran escala en las últimas décadas, y brinda crucial apoyo para esta labor. Tiene capacidad de analizar y procesar datos estandarizados y sin estandarizar; también puede extraer características y patrones, categorizarlos y almacenarlos para recuperarlos después, cuando el proceso lo amerite.

Profundizando un poco en la inteligencia artificial, Liebowitz [5] refiere las primeras ideas de mecanizar el pensamiento humano, promovidas por Aristóteles y Euclides. Por supuesto, para la época no se contaba con la tecnología que

potenciara tales ideas para convertirlas en una realidad. Fue entre los años 1956 y 1970 cuando las computadoras ayudaron a implementar gestores basados en reglas y sistemas expertos. Estas computadoras generaban lo que se conoce como razonamiento simbólico, a partir de la supervisión de ingenieros que aportaban el conocimiento y la validación de los resultados producidos. Este periodo marca la primera ola de la inteligencia artificial, plasmando en la realidad aquella idea de Aristóteles y Euclides de mecanizar el razonamiento humano.

Una rama de la inteligencia artificial es conocida como aprendizaje automático o mecanizado. Shwartz y David exponen su propósito e importancia de manera muy clara [17]. La complejidad de los problemas y la necesidad de adaptarse a un ambiente o situación particular, son dos aspectos que contribuyen a que un sistema o máquina aprenda a partir de la experiencia previa. Esta experiencia o conocimiento previo es un insumo fundamental para la toma de decisiones. Es necesario resaltar la importancia de este aprendizaje por medio de sistemas y máquinas, ya que existen tareas que al ser humano le tomarían mucho tiempo, asegurando además resultados precisos y sin errores. Algunos ejemplos para considerar son el análisis de grandes cantidades de datos, predicciones, búsquedas y extracción de información valiosa de manera automática, entre otros.

En las entrañas del aprendizaje automático existe una basta cantidad de tecnología y conceptos matemáticos del área de estadística, estos colaboran entre sí para hacer una realidad las predicciones. S. Clarke y L. Clarke exponen de manera acertada [18] la necesidad de crear predicciones. Explican que su propósito se debe a la necesidad de conocer ciertas circunstancias que podrían suceder, y tomar decisiones a partir de datos sólidos que indiquen con una gran probabilidad que pueden ser evitados o, por el contrario, ejecutar una acción de forma anticipada que favorezca determinado objetivo o estrategia, ya sea personal, grupal u organizacional.

Detrás del análisis predictivo como lo exponen V. McCarthy, M. McCarthy, Ceccucci y Halawi [19], existe un conjunto de modelos estadísticos avanzados y técnicas de aprendizaje automático o machine learning que se basan en datos históricos. Con el objetivo de que las predicciones sean efectivas, este insumo de datos históricos debe ser representativo. Para lograr esto, es necesario el proceso de limpieza, análisis y preparación de los datos, de forma que se elimine el *ruido* que pueda disminuir la efectividad de la solución desarrollada.

Los algoritmos son un conjunto de pasos que resuelven un problema. Este concepto también aplica en la analítica predictiva; de hecho, se dividen en algoritmos de aprendizaje supervisado y algoritmos de aprendizaje no supervisado [19]. En el primero se utilizan técnicas donde los datos que el sistema automatizado utiliza son correctos, etiquetados y validados por un observador, mientras que en el caso del segundo algoritmo, los datos no están etiquetados, y por lo tanto, el sistema debe encontrar patrones, categorías y características que le generen ese aprendizaje con la ausencia de alguien o algo que lo corrija, de tal manera que la información producida es nueva para el observador.

Otro concepto relevante para este trabajo, es la computación en la nube o cloud computing, el cual, de acuerdo con Amazon, uno de los pioneros en esta tecnología [20], consiste en aquella distribución de recursos tecnológicos que se usan bajo demanda, o sea, cada vez que el usuario lo requiera en el momento y lugar que así lo necesite. Para utilizarlo, es necesario que el usuario tenga conexión a internet y pague por su uso de acuerdo con la frecuencia que el proveedor establece. Entre esta distribución de recursos tecnológicos se pueden mencionar la capacidad de almacenamiento de datos, la capacidad informática y la disponibilidad de equipo de cómputo.

Una de las razones por las cuales en este trabajo se menciona la computación en la nube, obedece a los beneficios que trae para analizar los datos, prepararlos, proyectar predicciones y publicar la solución de manera fácil y rápida, de tal manera que genere resultados significativos en el corto plazo para la sociedad costarricense. Algunos de estos beneficios son: la agilidad con la que fácilmente se pueden utilizar herramientas tecnológicas que faciliten nuestra labor, el ahorro de costos al no tener que comprar equipo especializado ni contratar colaboradores que se encarguen del mantenimiento, control y soporte de la solución. Aunado a esto, se destaca también la elasticidad de la solución en la nube, ya que, de acuerdo con la demanda, así se comporta su capacidad y disponibilidad.

IV. Análisis de Resultados

Realizada la investigación teórica, se analizaron los datos del Poder Judicial para entender el comportamiento de los delitos y de esta manera establecer un modelo matemático predictivo que proyecte información sobre potenciales delitos en el próximo quinquenio en una región determinada, así como el incremento o decremento mensual durante este mismo periodo.

Los datos para el análisis van desde enero del 2015 hasta diciembre del 2019. Debido al comportamiento atípico que azotó al mundo, a inicios del 2020, se excluye este año del estudio para no generar anomalías en los resultados, pues los cierres totales y parciales que ha tenido Costa Rica tanto en lugares de recreación como en transporte desde marzo del 2020, pueden haber incidido en una disminución significativa en la cantidad de delitos cometidos en la nación, reduciéndolos casi a la mitad, en contraste con el año anterior. De la misma manera, esta investigación busca generar un modelo que pueda ser replicado en condiciones similares a las ocurridas antes de la pandemia, por ende, los resultados arrojados se dan con base en lo proyectado para un modo de vida sin restricciones sanitarias, pandemias ni otros eventos magnos que afecten el diario vivir de los costarricenses.

De acuerdo con los pasos explicados por Liebowitz [5] en el campo de la analítica de datos, y particularmente la ciencia de datos, se inició con la recolección, mediante la descarga de los datos disponibles en el sitio del Poder Judicial [3]. Los pasos

de procesamiento y gestión involucraron actividades de limpieza y cambio, eliminando datos incompletos, repetidos y que no aportan información relevante al estudio.

Una bondad de los análisis de datos es la posibilidad de generar predicciones con base en información histórica; además, sin importar cuál sea la metodología de predicción seleccionada, siempre es posible ejecutar los siguientes pasos: definición del proyecto, incluyendo lineamientos y objetivos; la exploración, donde se determina el método para recolectar información y su rango; la preparación de datos, en la cual estos se procesan para el estudio; construcción del modelo, donde se crean y evalúan las métricas; implementación, donde se aplican los resultados a los modelos; y finalmente, la gestión, donde se evalúan las mejoras necesarias para la evolución continua del modelo predictivo.

Aprovechando los datos abiertos brindados por el Poder Judicial, se utilizaron las Estadísticas Policiales comprendidas entre los años 2015-2019, para proyectar cómo el accionar de los delitos en distintas segmentaciones, como región, cantón, edad y género; así como evaluar la situación actual utilizando analítica descriptiva. Como producto final, se brinda una herramienta abierta en Tableau Public que queda abierto a consulta de la persona que lo desee al visitar el sitio web. Este análisis puede materializarse en acciones preventivas ejecutadas por las distintas entidades encargadas de la seguridad, desde Policía Municipal o Fuerza Pública, realizando contenciones o desplegando a sus unidades en lugares de mayor riesgo, aprovechando de mejor manera los recursos y el personal policial.

Parte esencial de la investigación es contar con datos que permitan un análisis apropiado. Para este estudio se empleó información de la sección de delitos de datos abiertos del Poder Judicial de Costa Rica sobre los actos delictivos ocurridos en el país entre 2015 y 2019. Con respecto a la segmentación, se aprovechan las variables que contienen estos datos para determinar la provincia, el cantón y el distritito donde se perpetró el delito, así como el tipo, el género de quien lo cometió, el rango de edad del presunto delincuente y el tipo de víctima. Se enriquece esta información con la segmentación utilizada por el MIDEPLAN para regionalizar los distritos del país y categorizarlos en urbano-rural, con el objetivo de realizar análisis macro más relevante; la fuente primaria de la toma de esta información es el Manual de Clasificación Geográfica con Fines Estadísticos de Costa Rica, desarrollado por el Instituto Nacional de Estadísticas y Censos (INEC) [21].

La información obtenida mediante la descarga de datos en línea del sitio web, se utiliza para crear la matriz de análisis y aplicar el estudio correctamente. Los datos pasan por un proceso de limpieza y preparación, para unirse de manera apropiada. Por ejemplo, mientras la información brindada por el INEC nombra un distrito como “La Fortuna”, el Poder Judicial lo menciona como “Fortuna”. Otros problemas que se encontraron fueron nombres de cantones tildados en algunas líneas y otras sin tildar; pronombres descritos incorrectamente en algunos casos. Se encontraron también algunos nombres extensos en su forma abreviada; entre otros. Este mismo patrón se presentó al revisar datos para los distritos. El mayor inconveniente se da porque el Poder Judicial realiza la carga de datos de diversos lugares y no aplica ningún filtro o limpieza antes ejecutarla. Este proceso se realiza en los cinco archivos CSV de las Estadísticas Policiales del Poder Judicial, uno por año. Estos se consolidan en un archivo maestro final, para poder unirlo con la información del INEC. Posterior a la descarga, limpieza y consolidación de las fuentes de datos, para facilitar esta homologación con respecto al nombre de los cantones y distritos, se agregan los códigos postales, para unir las fuentes con la menor cantidad de errores.

Con respecto a los datos geográficos brindados por el INEC, se tiene una fuente de datos que se encuentra en formato PDF, por lo que debe extraerse la información a un archivo tabular, de preferencia CSV. Esta labor se realizó con una herramienta web llamada I Love PDF [22], que permite convertir archivos de este formato a Excel. Como resultado, el archivo Excel posee varias hojas de cálculo que deben también pasar por un proceso de limpieza y consolidación en una sola fuente. Los errores más comunes en esta parte son espacios duplicados o antes de empezar o terminar un nombre, también se tienen filas con los nombres de columnas desalineados, afectando principalmente la línea inicial y final de cada hoja cálculo. Una vez realizado este proceso, se consolida en una sola hoja sin formato de ningún tipo para almacenarse en formato CSV.

Las bases de datos se unen en las herramientas seleccionadas para el análisis, utilizando las funciones predeterminadas para esta labor. En el mercado existe una cantidad considerable de formas para procesar, analizar y visualizar datos, para efectos de este análisis se seleccionaron varias herramientas que al complementarse permiten brindar una solución más acorde con la actualidad donde la mayor parte del esfuerzo es crear una solución visualmente atractiva para el consumidor: Microsoft Excel para la limpieza y consolidación de datos, Tableau Desktop, se utilizó para el análisis descriptivo y visualización de datos, Python para los análisis predictivos.

El análisis descriptivo se utiliza en estadística para describir lo que ha pasado, es decir, aquí se toman las variables categóricas y se realizan estudios según los distintos segmentos. Este análisis permite contestar interrogantes como el lugar donde se han cometido más delitos, qué diferencias hay entre estos, el género de quien lo cometió, su, edad y otros, así como una mezcla de diferentes variables para entender el comportamiento de la información. Este estudio es primordial para comprender de manera profunda la información con la cual se trabaja, además, permite entender la situación actual y tener indicios de cuáles áreas se desean investigar más a fondo.

También tenemos análisis prescriptivos y predictivos. El enfoque de esta investigación desarrolla el análisis predictivo para contestar preguntas como la propensión de que se cometa un delito en una región determinada. Para el análisis predictivo toma información histórica, aplicando modelos de estadística, siendo la más apta la regresión lineal, un modelo matemático

utilizado para aproximar la relación de dependencia entre una variable dependiente y variables independientes. La clave de utilizar modelos predictivos es comprender que se trata de una estimación y saber que no es exacta, ya que siempre existe un margen de error [23].

Tableau permite realizar una unión de las fuentes de datos, con un simple jalar y arrastrar los archivos y seleccionando la variable (o conjunto de ellas) común entre las dos bases de datos, en este caso, el campo seleccionado es el código distrital o código postal. El análisis descriptivo se realizó en esta herramienta, generando un tablero o dashboard analítico interactivo que permite evaluar los distintos tipos de segmentación, así como ver fácilmente los cambios históricos con tan solo aplicar filtros o arrastrar nuevas dimensiones a la vista. Este dashboard dinámico queda accesible a quien desee consultar en el sitio de Tableau Public [24]

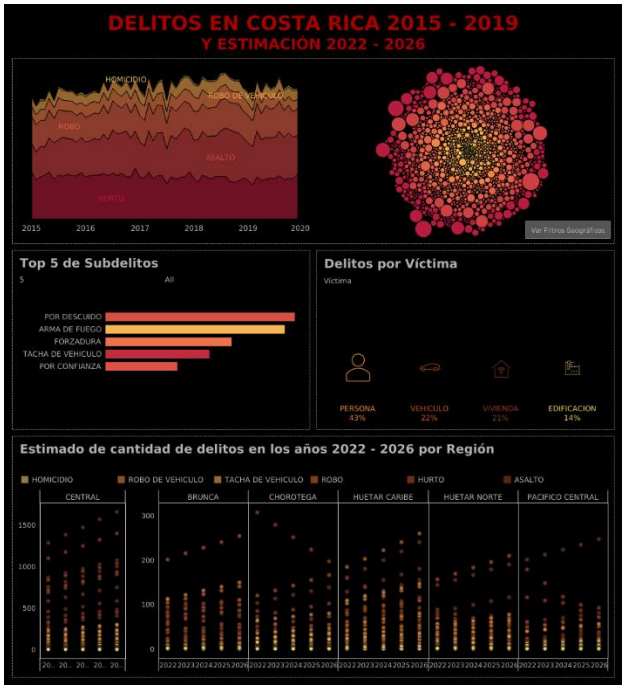

Figura 1. Delitos en Costa Rica 2015 – 2019
Fuente: Elaboración propia, 2021.

Otros datos del estudio se procesaron mediante Python; Este es uno de los lenguajes más utilizados por analistas y científicos de datos para el estudio de grandes cantidades de información. Mediante un programa creado específicamente para la predicción, se realiza un barrido de los datos del Excel. Para la creación del programa de análisis en Python se cuenta con una librería llamada sklearn [25], la cual contiene una gran cantidad de herramientas estadísticas, entre las cuales están la regresión y predicción (no es recomendable utilizarla para manipular datos u obtener resúmenes).

Los análisis de datos procesados mediante Python, se mueven a Tableau y se crean gráficos que pueden ser filtrados más fácilmente, permitiendo llevar la información a niveles más granulares y generando un diseño visualmente más atractivo.

Además, se utilizan dos funciones para la predicción en la herramienta Excel: slope e intercept. La función slope, que significa pendiente, calcula la pendiente de la regresión lineal cuando se le asignan dos puntos de datos normalmente establecidos en ejes X, Y. Luego se utiliza la función intercept (intercepción, según su traducción) para ajustar las funciones de predicción, podría decirse que es una extensión del diagrama hasta un punto futuro donde las variables se unen, lo que predice datos continuos hasta su intercepción.

En el análisis predictivo, se establece un coeficiente 0.65 o 65% como mínimo requerido para realizar proyecciones con una mayor solidez matemática, de tal manera que la ecuación para regresión lineal dada por y = mx + b genere predicciones más eficaces. Siendo *y* el número de delitos a predecir, *m*; el valor de la pendiente que muestra si la relación de los datos es positiva o negativa, y *b*; el valor del coeficiente de determinación que indica la eficacia de la expresión matemática para pronosticar resultados futuros de la variable *y*.

Los primeros resultados muestran un total de 292 315 delitos distintos a lo largo de cinco años, distribuidos casi de manera uniforme en cada año, con una media aproximada de 58 000 eventos por año a excepción del año 2018 donde se superan los 61 000 sucesos. Divididos en seis categorías principales, estos casi 300 000 delitos se distribuyen de la siguiente manera: 32% hurtos, 30% asaltos, 21% robos, 9% tacha de vehículos, 8% robo de vehículos y 1% homicidios. En cuanto a división geográfica provincial, un 38% de los hechos ocurrieron en San José, Alajuela con un 16%, seguido de Puntarenas con un 12%, Heredia y Limón con 9% cada uno, Guanacaste detrás con un 8% y Cartago con un 6%.

Las víctimas son en su mayoría personas, representando un 43% del total de los incidentes, mientras que vehículos y viviendas rondan el 21%, un 14% corresponde con edificaciones. Los victimarios son en su mayoría son del género masculino (67%), mayores de edad (86%), un 4% de los delitos fue cometido por menores de edad, y un 5% por adultos mayores.

Es necesario comprender la distribución de los datos para decidir cuáles potenciar para un análisis predictivo más profundo. Por ejemplo, debido a su alta dispersión, realizar un análisis predictivo para el siguiente homicidio no arrojará una cifra confiable.

Por otro lado, al enriquecer esta información con datos como por la regionalización, es factible señalar comportamientos específicos en ciertas zonas del país. Tomando como ejemplo Heredia, donde más de la mitad de su territorio está ubicado en la Región Huetar Norte (cantón de Sarapiquí) y el resto en la Región Central, se infiere que, por sus condiciones socioeconómicas, ambos espacios tienen comportamientos muy distintos, a pesar de pertenecer a la misma provincia. Un nivel granular por cantón o distrito, podría

disimular estas diferencias. Salvo el cantón central de San José y de Alajuela, se carece de información suficiente para realizar análisis predictivos significativos. Estos fenómenos se conocen como *outliers*, y pueden generar una interpretación errónea si se toman como normales dentro de la distribución de los datos.

En el análisis descriptivo por región, los hurtos son los delitos más comunes, a excepción de la Región Central y la Huetar Caribe, donde los asaltos predominan. Incluso, en las otras cuatro regiones, antes que los asaltos, los robos ocupan el el segundo lugar. Otra información destacable, es el lugar que ocupa la tacha de vehículos más alta: el Pacífico Central, siendo significativamente mayor que sus pares costeros; de hecho, el total general de tachas de vehículo es el 15%, mientras que en las otras regiones oscila entre el 4 y el 9%.

En la figura 2 se resume de manera visual la información descrita anteriormente, donde las barras muestran la cantidad de delitos, ordenados de mayor a menor cantidad, mientras que los colores representan cada una de las regiones y la injerencia proporcional que tienen en cada uno de estos incidentes.

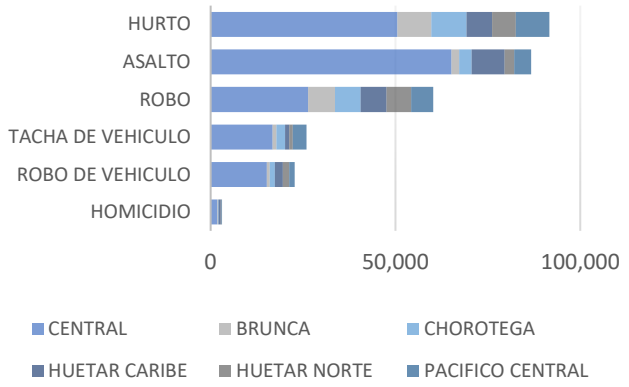

Figura 2. Delitos por región
Fuente: Elaboración propia, 2021.

Cuando se realiza el análisis temporal para comprender la evolución de los delitos a lo largo del tiempo y resaltar alguna diferencia en el comportamiento se destaca por región: en la Región Central se ilustra un aumento en los asaltos de cinco puntos porcentuales entre el 2015 y el 2019, subiendo del 34 al 39%. En la Brunca, en el 2015 los hurtos representaban la mitad de los delitos, decreciendo al 42% al final del quinquenio, mientras aumentaron los robos. En la Huetar Norte predominan los robos, seguidos por los hurtos, comportamiento que se mantiene de manera estable en todos los periodos, salvo en el año 2016, donde los robos disminuyen un 5%, dejando en primer lugar a los hurtos con un 38%, sin embargo, en el 2017 vuelve a su estado normal. El otro detalle que puede escapar al ojo un incremento en robo de vehículos, moviéndose de un 5% a un incremento paulatino, llegando a un 12% en el año 2018 y 11% en el 2019.

Por su parte, en la Región Chorotega se muestra un incremento en los asaltos, del 12 al 17% al finalizar el periodo en estudio, mientras que los hurtos disminuyen del 46 al 38%. La Región Huetar Caribe es uno de los casos más particulares, pues los delitos de asalto, robo y hurto empatan en 30% al empezar el periodo del análisis, poco a poco aumentan el porcentaje de asaltos hasta 38% y los robos y los hurtos disminuyen al 23%. Finalmente, en el Pacífico Central se muestra una tendencia a la baja en cuanto a los hurtos, que inician en un 42% y terminan en un 33%/ Los asaltos y robos de vehículos aumentan en oposición a esta distribución.

El análisis predictivo realizado en cada región pretende seguir la tendencia de comportamiento de cada una de estas regiones, apoyado principalmente en las características destacables del análisis descriptivo mencionado anteriormente. Los años para los que se proyecta este análisis corresponden al siguiente quinquenio: desde el 2022 al 2026.

La Tabla 1 muestra los coeficientes o porcentajes de confianza obtenidos del modelo de proyección, por delito y por región. Este valor se obtiene del resultado de la regresión lineal del modelo ejecutado en Python. Como se menciona anteriormente, se trabaja con un coeficiente mínimo de 0.65 para confiar en el modelo. Entre más alto el porcentaje, mayor es la probabilidad de que el modelo acierte en su propensión.

Delito	*Central*	*Brunca*	*Choro-tega*	*Huetar Caribe*	*Huetar Norte*	*Pac. Central*
Asalto	84.2%	82.0%	88.0%	82.5%	84.2%	86.6%
Homici-dio	96.2%	96.0%	98.5%	92.9%	92.2%	93.9%
Hurto	83.2%	77.7%	80.8%	86.0%	83.1%	86.8%
Robo	82.4%	84.1%	76.5%	89.4%	85.0%	85.1%
Robo de vehiculo	86.0%	86.2%	93.5%	82.9%	77.4%	84.8%
Tacha vehiculo	89.0%	95.5%	87.0%	97.0%	94.8%	89.0%

Tabla 1. Coeficiente de Confianza de Predicción por Región y Delito
Fuente: Elaboración propia, 2021

De las proyecciones se destacan las tendencias que llaman la atención, tanto basado en la información que nos demuestran los datos de los periodos 2015 al 2019, así como información que sea destacable considerar. De aquí podemos señalar que además de un análisis de regresión lineal, desarrollamos un estudio basado en la tendencia, como complemento al modelo. El enfoque se mantiene como se menciona en la tabla anterior, por región socioeconómica y por tipo de delito.

Así tenemos que la Región Central proyecta un aumento en los asaltos de 9 000 en el 2022 a 11 000 en el 2026, esto significa una ligera tendencia a la baja, comparada con el histórico, con una media aproximada de 12 000 asaltos al año, pero volviendo a subir con el tiempo; cabe destacar que hay una proyección de ligero aumento en los homicidios, de 114 a 163 del 2022 al 2026. Este ejemplo de los homicidios no es una alarma que nos demuestran los datos históricos, sin embargo, la proyección y su alto coeficiente lo destacan como punto de atención futura.

Para la Región Brunca se confirma una tendencia a la baja en los hurtos, con un porcentaje de confianza del 78%, lo cual

es bajo comparado con las otras estimaciones, sin embargo, entra dentro del margen definido. La tacha de vehículos se muestra como uno de los temas a destacar, con un aumento de 124 sucesos a 168 en los rangos anuales mencionados.

En la Región Huetar Norte se proyecta una tendencia al alza en robo de vehículos, el cual se confirma con la regresión lineal, pasando de aproximadamente 400 a casi 500 en los siguientes 5 años. Los asaltos tienden a aumentar, mientras que los hurtos tienden a la baja; en general los valores de aumento y disminución no generan un hallazgo significativo.

En la Región Chorotega la tendencia demostrada en el quinquenio anterior se proyecta en el siguiente: aumento en los asaltos y disminución en los hurtos. Los asaltos aumentan de unos 132 en 2022 a 181 en 2026, mientras que los hurtos pasan de 639 a 440 en el mismo periodo. Hay dos factores interesantes en la proyección que no precisamente se demuestran en los datos históricos: el primero es un aumento en los homicidios de un 30% en el quinquenio siguiente, el segundo corresponde a un aumento del 25% en el robo de vehículos.

En la Región Huetar Caribe se proyecta un aumento en los asaltos, pero una disminución en los robos y los hurtos. Los asaltos muestran un aumento del 30% en los siguientes 5 años, mientras que los hurtos disminuyen la misma cantidad, por otro lado, los robos se proyectan en 441 en 2022 a 392 en 2026, una baja paulatina. Uno de los delitos con mayor coeficiente que arrojó este análisis, es un incremento en la tacha de vehículos de un 20% aproximadamente para los cinco años siguientes.

Finalmente, la Región Pacífico Central sigue tendiendo al alza en esta materia, proyectando un incremento de 12 a 17 carros afectados por tacha. Los homicidios también son un factor de preocupación donde se proyecta un aumento de más del 40% en el siguiente periodo; siendo un nivel de confianza por encima del 0.9. Los robos y asaltos también tienden a subir. En general, esta región se perfila con un aumento en su peligrosidad para el siguiente quinquenio.

Además del análisis de macro delitos por región, se realizó un análisis temporal para comprender cómo pueden comportarse los delitos por mes en los años siguientes. Según los datos registrados y aprovechando la misma técnica matemática, se establece un incremento mensual de delitos de asalto en el país para el próximo quinquenio en enero, febrero, abril, junio, noviembre y diciembre. Por el contrario, se proyecta una disminución en la cantidad de delitos para los meses no incluidos en la lista anterior. En la figura 3 se muestran las predicciones donde se responde a las interrogantes y a los detalles descritos previamente. Cabe aclarar que este análisis mensual proyecta una relación predictiva de la cantidad de asaltos futuros con un porcentaje de precisión de un 52.8%.

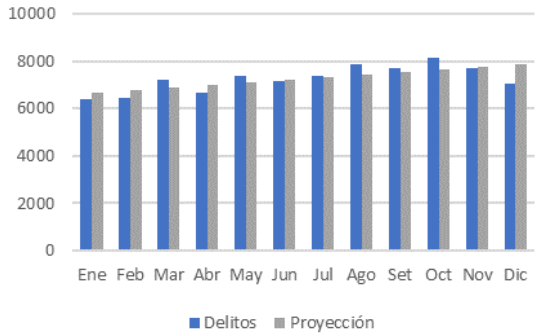

Figura 3. Proyección de delitos mensuales
Fuente: Elaboración propia, 2021.

En este análisis destacamos tres meses que vale la pena acotar. En diciembre se proyecta un incremento cercano al 10% de los delitos denunciados en esta época, siendo el de mayor incremento en las contravenciones esperadas para el siguiente quinquenio. Mas no todo son malas noticias, en los meses de octubre y agosto se espera una reducción de los delitos en 6%, comparado con el periodo estudiado, siendo éste un punto importante a destacar del análisis resultante del modelo de proyección mensual.

Con respecto a las limitaciones del estudio, una de las interrogantes no pudo ser contestada con los datos disponibles: proyectar el margen de edad más propenso de cometer un delito en los próximos cinco años. Sin embargo, los datos solamente ofrecen las siguientes categorías: mayor de edad, adulto mayor, menor de edad y desconocido. Después del análisis hecho y haciendo la respectiva correlación de estas categorías de edades con respecto al número de delitos, se concluye que la tendencia de las víctimas es significativamente alta hacia los mayores de edad. En el caso de los hurtos del año 2015 se tiene 16 349 víctimas mayores de edad, mientras que los menores de edad suman 1 305 y las personas adultas mayores 986. Un comportamiento similar surge en los siguientes años analizados.

Dada esta tendencia, el análisis de regresión lineal genera una relación débil y dispersa lo cual afecta de forma importante la efectividad de la predicción. Aun así, realizando la predicción con un coeficiente de determinación del 62.1% como el obtenido en el análisis, el aporte que generan los resultados no es significativo, ya que sigue la tendencia de proyectar a los mayores de edad como los más propensos a sufrir delitos. Esta predicción generaría un mayor valor si los datos disponibles tuvieran edades en números o rangos de edad más detallados; por ejemplo, separados por años, quinquenios o décadas.

V. Conclusiones y Recomendaciones

Las instituciones gubernamentales de Costa Rica deben adoptar mejores prácticas de análisis de transparencia, como publicar constantemente los datos relevantes a su institución. Según convenios internacionales, los ciudadanos deberían tener acceso a información que incluye desde salarios de funcionarios públicos hasta ejecución presupuestaria, planes de trabajo, entre otros. En el caso del Poder Judicial, intenta cumplir con estos

lineamientos, al tener disponible esa información y otras estadísticas relevantes, como datos de feminicidios, violencia intrafamiliar y estadísticas policiales, sin embargo, sin una comunicación adecuada, es como gritar al vacío. La primera recomendación es generar el hábito en los entes gubernamentales de publicar y promocionar la información, para fomentar la transparencia estatal y animar la participación de sociedad civil en el ejercicio de fiscalizar las instituciones.

Se entiende que análisis de datos pasa por muchas etapas de recolección y limpieza, a pesar de que los productos brindados por el Poder Judicial y el INEC son útiles y completos, siempre es necesaria una limpieza rigurosa para conectar las fuentes. Este primer paso es fundamental en todo proceso de análisis, porque además tiene como efecto secundario una cercanía entre el analista y el dato, comprendiendo cómo interacciona cada uno de los registros entre sí. Luego de unir los datos, es imperativo verificar que no haya duplicidad en la información. Las herramientas de análisis de datos facilitan esta labor, de ahí que las seleccionadas para este ejercicio hayan sido Excel, Tableau y Python, pues su uso intuitivo agiliza esta etapa.

El modelo predictivo utiliza una mezcla de análisis de tendencia con regresión lineal, pues es recomendado utilizar más de una técnica de propensión. Para interpretar los resultados de manera más relevante, el estudio inicial de lo que muestran los datos del quinquenio base para extraer las conclusiones que pueden ser más relevantes basado en el comportamiento de las tendencias pasadas. Pero no es lo único que se toma, datos con alto margen de precisión o confianza, o cambios importantes en los datos, también se destacan como parte de los resultados del modelo.

Cuando se formula la investigación, se busca aprovechar muchos de los datos de segmentación que incluye la base, sin embargo, se busca ampliar la base de datos, enriqueciendo la información utilizando regionalización, para brindar detalles más profundos sobre el comportamiento de localidades con condiciones socioeconómicas similares, aprovechando el gran estudio que ha realizado en materia entidades como el MIDEPLAN y el INEC, ya que la georreferenciación de división política tradicional, compuesta por provincia ,cantón y distrito, no siempre es la más exacta para analizar algún tipo de comportamiento social en una zona, dado a los contrastes que tienen las provincias y cantones, según el área donde se encuentren.

Por medio de los análisis descriptivos se demostró cuál es la realidad nacional en cuanto a crimen, según las denuncias impuestas en el OIJ. Se observa que la Región Central, especialmente los cantones de San José y Alajuela centro encabezan las listas de zonas inseguras y que el comportamiento de esta zona difiere mucho a las regiones más alejadas el Valle Central. Los fenómenos socioeconómicos de cada región, la cantidad de población y viviendas son factores determinantes cuando se ve esta información, por lo que no es recomendable comparar regiones entre sí, pues cada una tiene su propia realidad. Uno de los resultados ilustrados que llama la atención es la Región Pacífico Central y la tacha de vehículos, cuya incidencia supera la media, tener claro esta predicción permite un accionar preventivo en cuanto a seguridad de los carros en esta zona. Otro tema que cabe destacar es que los homicidios no superan el 1% de los delitos en Costa Rica, lo cual es un buen augurio, sin embargo, La región Caribe es un caso especial, y duplica la cantidad de homicidios nacionales, desde una perspectiva porcentual.

Por ende, viendo los datos históricos se puede pensar que existe una alta probabilidad de asaltos en la Región Central, de tacha de vehículo en el Pacífico y homicidios en la Huetar Caribe, en comparación porcentual de los delitos entre las regiones. La Región Chorotega junto con la Brunca evidencian las proyecciones de delitos más bajas.

El análisis predictivo realizado buscó tomar cada una de las características destacadas de cada región y verificar cuál será el comportamiento en el siguiente quinquenio. Además, se encontraron otros datos interesantes, donde se destaca una tendencia al alza en homicidios en varias regiones del país, con altos porcentajes de fiabilidad. Sin embargo, es una estimación. Es recomendable seguir trabajando con este modelo para su perfeccionamiento.

Durante el análisis hubo preguntas que no lograron contestarse debido a la alta dispersión de los datos como, por ejemplo, la información por distrito. Otros porque la segmentación no genera un valor significativo, como género o edad. Una de las recomendaciones para el Poder Judicial es recopilar información relevante sobre el victimario y la víctima, como, por ejemplo, el género de ambas partes. Una mejora que puede generar mucho valor es incluir la edad exacta de quien comete el delito, y también, de ser posible, la víctima.

En general, durante esta investigación se logró cumplir los objetivos propuestos, pues en un periodo corto se desarrolló un modelo de datos predictivo que permite visibilizar los delitos por región, y un producto de análisis de datos descriptivo que ilustra la situación de los delitos en Costa Rica, accesible para cualquier persona que quiera investigar el tema.

VI. Trabajos Futuros

Dados los resultados obtenidos en la investigación y el desarrollo del trabajo práctico, se requiere implementar un mecanismo de alimentación constante con datos actualizados trimestralmente con el fin de mantener activa y actualizada la solución propuesta con el modelo, brindando información valiosa a la sociedad costarricense. El mecanismo automatizado debe descargar los datos generados por el Poder Judicial, limpiarlos y almacenarlos en un repositorio en la nube como Google Drive en formato CSV; donde sea accedido por la herramienta predictiva, y luego de la aplicación de técnicas de regresión lineal, ser proyectados mediante un enlace de acceso público hacia las personas, en Tableau.

Aunado a eso, se sugiere involucrar a científicos de datos y estadísticos para el desarrollo de un análisis matemático más profundo con el fin de evaluar mejoras en la precisión de las predicciones desarrolladas y futuras. Estos perfiles ayudan a

fortalecer el modelo predictivo mediante pruebas más rigurosas basadas en fundamentos matemáticos más avanzados. Este trabajo permitirá exponer con claridad el esfuerzo requerido para contestar a preguntas como el incremento o decremento mensual de cada delito por región de Costa Rica para el próximo quinquenio.

Adicionalmente, se sugiere extraer más información de los datos disponibles de forma pública por las diferentes instituciones gubernamentales, de forma que se muestren las acciones que otras instituciones en Costa Rica están ejecutando en materia de seguridad, y posibles integraciones de datos institucionales que fortalezcan las soluciones preventivas mediante herramientas tecnológicas con un enfoque predictivo.

REFERENCIAS

[1] J. y. o. Madrigal Pana, "RESULTADOS DE LA ENCUESTA ACTUALIDADES 2012," 2012. [En línea]. Disponible en: http://www.ucr.ac.cr/medios/documentos/2012/UCR-ESTADISTICA-ACTUALIDADES-2012.pdf. [Accedido: 21 Agosto 2021].

[2] "Visualizador," Programa Estado de la Nación, [En línea]. Disponible en: http://estadisticas.estadonacion.or.cr/visualizador. [Accedido: 19 Agosto 2021].

[3] Poder Judicial, "Datos Abiertos," Poder Judicial, [En línea]. Disponible en: https://pj.poder-judicial.go.cr/index.php/rendicion-de-cuentas/datos-abiertos. [Accedido: 27 Julio 2021].

[4] "Comunicados," [En línea]. Disponible en: https://www.colegiotopografoscr.com/comunicados/2018/creacioncanton.pdf. [Accedido: 20 Agosto 2021].

[5] J. Liebowitz, Data Analytics and AI, Taylor & Francis Hroup.

[6] O. D. Handbook, "¿Qué son los datos abiertos?," [En línea]. Disponible en: http://opendatahandbook.org/guide/es/what-is-open-data/. [Accedido: 21 Agosto 2021].

[7] ODC, "Open Data Charter," [En línea]. Disponible en: https://opendatacharter.net/. [Accedido: 21 Agosto 2021].

[8] A. G. Zúñiga, *Adopción de la Carta Internacional de Datos Abiertos,* San José, 2016.

[9] CEPAL, "América Latina y el Caribe: Países que cuentan con Ley de Acceso a la Información Pública y año de promulgación," 26 Noviembre 2018. [En línea]. Disponible en: https://observatoriop10.cepal.org/es/recursos/america-latina-caribe-paises-que-cuentan-ley-acceso-la-informacion-publica-ano. [Accedido: 21 Agosto 2021].

[10] Organismo de Investigación Judicial, "Manual de Usuario: Sistema de Estadísticas del OIJ," 2018. [En línea]. Disponible en: https://pjenlinea3.poder-judicial.go.cr/estadisticasoij/Image/Manual%20de%20Usuario.pdf. [Accedido: 21 Agosto 2021].

[11] C. Mora Izaguirre, M. Solano Chaves, J. Hernández Murillo, I. Rodríguez González y K. Hernández Hernández, "INFORME DE ENCUESTA: PERCEPCIÓN SOBRE LA SEGURIDAD EN COSTA RICA," Noviembre 2020. [En línea]. Disponible en: https://drive.google.com/file/d/1hll2ejyl4V_C7qQ4ldPv0FFZduCKcar0/view. [Accedido: 21 Agosto 2021].

[12] P. Trujillo Álvarez, "Violencia en Centroamérica: reflexiones sobre causas y consecuencias," 2017. [En línea]. Disponible en: https://reshistorica.journals.umcs.pl/al/article/view/5411. [Accedido: 18 Setiembre 2021].

[13] M. Sánchez Alvarado, "Reconstrucción de la Política de Prevención del delito en Costa Rica," Junio 2017. [En línea]. Disponible en: http://revistafacso.ucentral.cl/index.php/rumbos/article/view/26/21. [Accedido: 20 Agosto 2021].

[14] H. Villalobos Fonseca, "Revista de Relaciones Internacionales, Estrategia y Seguridad," Junio 2020. [En línea]. Disponible en: http://www.scielo.org.co/pdf/ries/v15n1/1909-3063-ries-15-01-79.pdf. [Accedido: 20 Agosto 2021].

[15] F. Arteaga, 24 Mayo 2018. [En línea]. Disponible en: https://www.almendron.com/tribuna/wp-content/uploads/2018/05/dt12-2018-arteaga-cuarta-revolucion-industrial-enfoque-seguridad-nacional.pdf. [Accedido: 20 Agosto 2021].

[16] J. M. Rábade Roca, "La Innovación Policial en la Ciudad del Siglo XXI," Enero 2018. [En línea]. Disponible en: https://www.ciudades-creativas.com/proceedings/6ccc/proceedings-6ccc_040.pdf. [Accedido: 19 Agosto 2021].

[17] S. Shalev-Shwartz y S. Ben-David, Understanding Machine Learning from Theory to Algorithms, Cambridge University Press, 2014.

[18] B. S. Clarke y J. L. Clarke, Predictive Statistics Analysis and Inference beyond Models, Cambridge University Press, 2018.

[19] R. V. McCarthy, M. M. McCarthy, W. Ceccucci y L. Halawi, Applying Predictive Analytics Finding Value in Data, Springer, 2019.

[20] Amazon, "¿Qué es la informática en la nube?," Amazon, [En línea]. Disponible en: https://aws.amazon.com/es/what-is-cloud-computing/. [Accedido: 22 Agosto 2021].

[21] Instituto Nacional de Estadísticas y Censos (INEC), "Manual de Clasificación Geográfica con Fines Estadísticos de Costa Rica," INEC, San José, Costa Rica, 2016.

[22] I love PDF, "Convierte PDF a EXCEL," [En línea]. Disponible en: https://www.ilovepdf.com/es/pdf_a_excel. [Accedido: 13 Agosto 2021].

[23] M. F. Triola, "Intervalos de confianza," de *Estadística*, Pearson Education, Inc, 2006, pp. 320-320.

[24] N. Aviles, F. Montero y D. Coto, "Delitos en Costa Rica 2015 - 2019 y Estimación 2022 - 2026," Tableau Public, San José, Costa Rica, 2021.

[25] scikit-learn, "scikit-learn," scikit-learn.org, [En línea]. Disponible en: https://scikit-learn.org/stable/index.html. [Accedido: 9 Setiembre 2021].