



**UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL**

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

**TEMA:**

**Modelo predictivo que analice cuales son las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y a partir de esta información predecir qué sectores en desarrollo tendrán el mismo nivel de delincuencia**

**AUTOR:**

**Flores Asinc Miguel Alfonso**

**Trabajo de Titulación previo a la obtención del título de  
INGENIERO EN SISTEMAS COMPUTACIONALES**

**TUTOR:**

**Ing. Miranda Rodríguez Marcos Xavier**

**Guayaquil, Ecuador**

**16 de febrero del 2023**



UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

## **CERTIFICACIÓN**

Certificamos que el presente trabajo de titulación, fue realizado en su totalidad por **Flores Asinc Miguel Alfonso**, como requerimiento para la obtención del título de **Ingeniero en Sistemas Computacionales**.

### **TUTOR**

f. \_\_\_\_\_  
**Ing. Miranda Rodríguez Marcos Xavier**

### **DIRECTOR DE LA CARRERA**

f. \_\_\_\_\_  
**Ing. Camacho Coronel Ana Isabel**

**Guayaquil, a los 16 días del mes de febrero del año 2023**



UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

## **DECLARACIÓN DE RESPONSABILIDAD**

Yo, **Flores Asinc Miguel Alfonso**

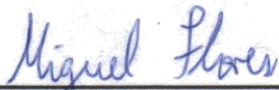
### **DECLARO QUE:**

El Trabajo de Titulación, **Modelo predictivo que analice cuales son las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y a partir de esta información predecir qué sectores en desarrollo tendrán el mismo nivel de delincuencia** previo a la obtención del título de **Ingeniero en Sistemas Computacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

**Guayaquil, a los 16 días del mes de febrero del año 2023**

**EL AUTOR**

f.   
\_\_\_\_\_  
**Flores Asinc Miguel Alfonso**



UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

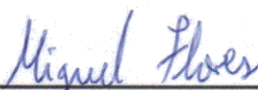
## **AUTORIZACIÓN**

Yo, **Flores Asinc Miguel Alfonso**

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, **Modelo predictivo que analice cuales son las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y a partir de esta información predecir qué sectores en desarrollo tendrán el mismo nivel de delincuencia** previo a la obtención del título de **Ingeniero en Sistemas Computacionales**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

**Guayaquil, a los 16 días del mes de febrero del año 2023**

**EL AUTOR:**

f.   
**Flores Asinc Miguel Alfonso**



UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERIA

CARRERA DE INGENIERIA EN SISTEMAS COMPUTACIONALES

## REPORTE URKUND

← BACK TO ANALYSIS OVERVIEW

RECYCLE DOWN HELP | PROFILE ▾

|                                 |                                       |            |
|---------------------------------|---------------------------------------|------------|
| SUBMITTER                       | FILE                                  | SIMILARITY |
| MARCOS XAVIER MIRANDA RODRIGUEZ | Trabajo_titulacion_Miguel_Flores.docx | 1 %        |



Firmado electrónicamente por:  
MARCOS XAVIER  
MIRANDA RODRIGUEZ

f. \_\_\_\_\_

**Ing. Miranda Rodríguez Marcos Xavier**

## TUTOR

f. \_\_\_\_\_

**Ing. Miranda Rodríguez Marcos Xavier**

## **AGRADECIMIENTO**

Mis más sinceros agradecimientos a mi familia que siempre me apoyo y ayudo en todo momento que necesite, mis padres y hermanos que siempre me preguntaban qué tal me iba, mis amigos con los que tuvimos buenos y malos momentos, momentos divertidos y estresantes y finalmente un agradecimiento a todos los docentes que me compartieron su conocimiento que me ayudo a poder realizar todo tipo de trabajos.

## **DEDICATORIA**

Le deseo dedicar mi proyecto de titulación a mis padres, mis hermanos que no se encuentran conmigo en el país que nunca paraban de apoyarme y darme de su cariño, les agradezco de todo corazón, a mis tíos que siempre me brindaban su apoyo, a mis amigos los cuales me ayudaban a entender cosas que no comprendían, este trabajo de titulación es para todos ustedes, gracias.



**UNIVERSIDAD CATÓLICA  
DE SANTIAGO DE GUAYAQUIL  
FACULTAD DE INGENIERÍA  
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

**TRIBUNAL DE SUSTENTACIÓN**

f. \_\_\_\_\_

**ING. ANA ISABEL CAMACHO CORONEL**  
DIRECTOR DE CARRERA

f. \_\_\_\_\_

**ING. ROBERTO GARCÍA SÁNCHEZ**  
DOCENTE DE LA CARRERA

f. \_\_\_\_\_

**ING. VICENTE ADOLFO GALLARDO POSLIGUA**  
OPONENTE



# ÍNDICE

|   |      |
|---|------|
| RESUMEN.....                                | XVI  |
| ABSTRACT .....                              | XVII |
| INTRODUCCIÓN.....                           | 2    |
| CAPÍTULO I.....                             | 6    |
| EL PROBLEMA .....                           | 6    |
| PLANTEAMIENTO DEL PROBLEMA .....            | 6    |
| Ubicación del Problema en un Contexto ..... | 6    |
| Causas y Consecuencias del Problema .....   | 6    |
| Delimitación del Problema.....              | 8    |
| Formulación del Problema .....              | 9    |
| OBJETIVOS.....                              | 11   |
| OBJETIVO GENERAL.....                       | 11   |
| OBJETIVOS ESPECÍFICOS.....                  | 11   |
| ALCANCES DEL PROBLEMA .....                 | 11   |
| JUSTIFICACIÓN E IMPORTANCIA .....           | 12   |
| HIPÓTESIS O PREGUNTA DE INVESTIGACIÓN ..... | 12   |
| VARIABLES DE LA INVESTIGACIÓN .....         | 13   |
| CAPÍTULO II.....                            | 14   |
| MARCO TEÓRICO .....                         | 14   |
| MODELOS PREDICTIVOS .....                   | 14   |
| MINERÍA DE DATOS.....                       | 15   |
| TÉCNICAS DE MINERÍA DE DATOS .....          | 17   |

|   |    |
|---|----|
| Árbol de decisión .....                               | 17 |
| Red neuronal .....                                    | 18 |
| Modelado estadístico.....                             | 19 |
| Reglas de asociación.....                             | 19 |
| Agrupamiento (Clustering) .....                       | 20 |
| Algoritmo genético .....                              | 20 |
| Regresión lineal.....                                 | 21 |
| Redes bayesianas .....                                | 21 |
| METODOLOGÍAS DE MINERÍA DE DATOS .....                | 22 |
| KDD (Knowledge Discovery in Databases).....           | 22 |
| SEMMA (Sample Explore Modify Model Assess).....       | 23 |
| CRISP-DM.....   | 24 |
| COMPARACION ENTRE KDD, SEMMA Y CRISP-DM.....          | 26 |
| LENGUAJES DE PROGRAMACION PARA MINERIA DE DATOS ..... | 26 |
| Python .....  | 26 |
| Características de Python.....                        | 29 |
| CAPÍTULO III.....                                     | 30 |
| METODOLOGÍA DE LA INVESTIGACIÓN .....                 | 30 |
| TIPO DE INVESTIGACIÓN .....                           | 30 |
| METODOLOGIA DE MINERIA DE DATOS.....                  | 31 |
| Selección de datos .....                              | 31 |
| Preprocesamiento y limpieza.....                      | 32 |
| Transformación.....                                   | 32 |

|  |           |
|--|-----------|
| Minería de datos .....                                   | 33        |
| Interpretación.....                                      | 33        |
| POBLACIÓN Y MUESTRA .....                                | 33        |
| <b>EL TAMAÑO DE LA MUESTRA.....</b>                      | <b>34</b> |
| CAPÍTULO IV .....  | 36        |
| PROPUESTA TECNOLÓGICA.....                               | 36        |
| HERRAMIENTAS TECNOLÓGICAS UTILIZADAS .....               | 36        |
| Herramienta de desarrollo .....                          | 36        |
| Lenguaje de desarrollo .....                             | 38        |
| Metodologías de minería de datos.....                    | 39        |
| Técnica de minería de datos.....                         | 40        |
| DESARROLLO DEL DISEÑO DE MODELO PREDICTIVO .....         | 41        |
| Diagrama del modelo predictivo usando KDD .....          | 41        |
| Obtención de Dataframe.....                              | 42        |
| Variables que determinan a un sector como peligroso..... | 44        |
| Preprocesamiento y limpieza de datos .....               | 48        |
| Técnica de minería de datos .....                        | 52        |
| Regresión lineal.....                                    | 53        |
| Random Forest.....                                       | 58        |
| Soporte de vectores.....                                 | 61        |
| Zonas con mayor índice de peligrosidad.....              | 63        |
| Determinar qué zonas se volverán más peligrosas.....     | 64        |
| Comparación de rendimiento de los modelos.....           | 66        |

|  |    |
|--|----|
| ANÁLISIS DE LAS PREDICCIONES OBTENIDAS ..... | 69 |
| CONCLUSIONES .....                           | 72 |
| RECOMENDACIONES.....                         | 75 |
| REFERENCIAS .....                            | 76 |
| ANEXOS .....                                 | 82 |

## INDICE DE TABLAS

|                      |            |
|----------------------|------------|
| <b>Tabla 1</b> ..... | <b>26</b>  |
| <b>Tabla 2</b> ..... | <b>37</b>  |
| <b>Tabla 3</b> ..... | <b>38</b>  |
| <b>Tabla 4</b> ..... | <b>39</b>  |
| <b>Tabla 5</b> ..... | <b>40</b>  |
| <b>Tabla 6</b> ..... | <b>45</b>  |
| <b>Tabla 7</b> ..... | <b>46</b>  |
| <b>Tabla 8</b> ..... | <b>467</b> |

## INDICE DE FIGURAS

|                        |    |
|------------------------|----|
| <b>Figura 1</b> .....  | 23 |
| <b>Figura 2</b> .....  | 25 |
| <b>Figura 3</b> .....  | 41 |
| <b>Figura 4</b> .....  | 43 |
| <b>Figura 5</b> .....  | 49 |
| <b>Figura 6</b> .....  | 50 |
| <b>Figura 7</b> .....  | 50 |
| <b>Figura 8</b> .....  | 51 |
| <b>Figura 9</b> .....  | 51 |
| <b>Figura 10</b> ..... | 54 |
| <b>Figura 11</b> ..... | 55 |
| <b>Figura 12</b> ..... | 55 |
| <b>Figura 13</b> ..... | 56 |
| <b>Figura 14</b> ..... | 56 |
| <b>Figura 15</b> ..... | 57 |
| <b>Figura 16</b> ..... | 58 |
| <b>Figura 17</b> ..... | 58 |
| <b>Figura 18</b> ..... | 59 |
| <b>Figura 19</b> ..... | 59 |
| <b>Figura 20</b> ..... | 60 |
| <b>Figura 21</b> ..... | 60 |

|                        |    |
|------------------------|----|
| <b>Figura 22</b> ..... | 61 |
| <b>Figura 23</b> ..... | 62 |
| <b>Figura 24</b> ..... | 62 |
| <b>Figura 25</b> ..... | 63 |
| <b>Figura 26</b> ..... | 64 |
| <b>Figura 27</b> ..... | 65 |
| <b>Figura 28</b> ..... | 65 |
| <b>Figura 29</b> ..... | 69 |
| <b>Figura 30</b> ..... | 70 |

## RESUMEN

En el presente proyecto de investigación contiene el diseño de un modelo predictivo que analice las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y con esta información predecir qué sectores en desarrollo tendrán un aumento de delincuencia usando la minería de datos y algoritmos de aprendizaje de máquina. Este modelado se lo hará con el uso de herramientas de programación con las que se podrá limpiar los datos, procesarlos y aplicando aprendizaje supervisado junto con algoritmos de regresión y random forest y así de esta manera entrenar el modelo y poder predecir valores futuros de la delincuencia en las diferentes zonas de Guayaquil y poder ser visualizados en una herramienta de visualización de datos. Todo el proceso desde la recopilación de datos hasta el análisis de los resultados obtenidos forma parte de las etapas de la metodología de minería de datos KDD o Knowledge Discovery in Databases. Los objetivos del proyecto de investigación constan de recopilar y analizar variables que describan a un sector como peligroso, diseño propio del modelo predictivo y la valoración de las predicciones obtenidas tras la creación del modelo. Por último, el proyecto concluye con la visualización de los algoritmos de regresión en una herramienta de visualización que presente las zonas que puedan convertirse en peligrosas en un futuro.

**Palabras Clave:** *Minería de datos, Aprendizaje autónomo, Algoritmos de regresión, Modelo predictivo, KDD*



## **ABSTRACT**

This research project contains the design of a predictive model that analyzes the areas with the highest index of danger in the city of Guayaquil and with this information predict which developing sectors will have an increase in crime using data mining and machine learning algorithms. This modeling will be done with the use of programming tools that will be used to clean the data, process it and apply supervised learning algorithms like linear regression and random forest to train the model and predict future values of crime in different areas of Guayaquil and be able to be visualized in a data visualization software. The whole process from data collection to the analysis of the results obtained is part of the stages of the KDD or Knowledge Discovery in Databases (KDD) data mining methodology. The objectives of the research project consist of collecting and analyzing variables that describe a sector as dangerous, the design of the predictive model itself and the assessment of the predictions obtained after the creation of the model. Finally, the project concludes with the visualization of the regression algorithms in a visualization software that presents the areas that may become dangerous in the future.

***Key words:*** *Data Mining, Machine Learning, Regression Algorithms, Predictive Models, KDD*

## INTRODUCCIÓN

Los países en el mundo que son considerados extremadamente peligrosos no son países que se encuentran en el continente americano, sino que estos destacan en países pertenecientes al continente africano como lo son Afganistán, Yemen, Siria, Libia, Malí, Somalia, Sudán del Sur y la República Centroafricana, junto con partes de Mozambique, Nigeria, la República Democrática del Congo, Ucrania, Pakistán, Irak y Egipto. que han sido clasificados como más peligrosos, pero no son categorizados con el mismo tipo de peligrosidad que países de Latinoamérica, ya que el índice de peligrosidad no depende tanto de la delincuencia, sino que depende de confrontaciones entre naciones mayormente por razones religiosas, por lo cual según Internacional SOS requieren un nivel diferente de contingencia. (Coffey, 2022)

Según Gerardo Lissardy, periodista de la cadena BBC News Mundo la región latinoamericana es considerada como una de las zonas más violenta del mundo. El autor resalta que según un informe publicado por las naciones unidas como en todos los países latinoamericanos todas las semanas se reportan homicidios representando un 37% de las muertes de todo el mundo, la cual representa una población de alrededor un 8%. La razón a que se deba tanta muerte es mayormente al crimen organizado, armas e impunidad, pero eso no quiere decir que solo en Latinoamérica existe el crimen organizado, la diferencia es que el continente americano es el único en el mundo que es productor del negocio del tráfico de cocaína, negocio que en que diferentes

carteles de diferentes países quieren dominar poniendo en peligro a todos los residentes de dichos países. (Lissardy, 2019)

Cuando se piensa en Latinoamérica, se piensa en ambientes tropicales, la flora, la fauna, las civilizaciones indígenas, la Amazonía, pero lamentablemente una palabra que a la que se asocia también es la delincuencia. Hace una década, cuando se hablaba de países peligrosos, los primeros países que a las personas pensaban eran mayormente países en que el negocio de tráfico de droga reinaba como México, Colombia o Brasil, pero ahora la red de narcotráfico ya se ha extendido de una manera en la que todos los países de Latinoamérica contienen grupos criminales pertenecientes carteles narco delictivos que se pelean por poder, lo cual ha dado como consecuencia un Centro América y Sudamérica inseguro.

El incremento de la delincuencia en Ecuador no se limita solamente a las ciudades más grandes del país como Guayaquil, Quito o Cuenca, sino que este afecta a toda las personas en una escala aun mayor, tanto así que en la provincia de Esmeraldas, las muertes han aumentado en 149 de lo que va el año en comparación al año 2021, y si se compara con el año 2015, las muertes promedios de ese año eran alrededor de 70, lo cual muestra un aumento del 100% en comparación al año 2021, lo cual indica un aumento drástico en el nivel delincencial que sufre el Ecuador.

Ahora centrándose en Ecuador, un país que al igual que otros países latinoamericanos, también presenta delincuencia que es consecuencia de las peleas por parte de grupos narcotraficantes y esta delincuencia no se centra solamente en las ciudades grandes como Guayaquil, sino que también está presente en otras provincias menos pobladas como por ejemplo Esmeraldas

que a pesar de ser una provincia de menor tamaño presenta un alto índice de delincuencia y esta razón se debe mucho al tener un puerto el cual es una vía directa para los carteles a exportar droga a Estados Unidos y Europa. (Pichel, 2021)

A pesar que en la actualidad existan ciudades y provincias igual o hasta más peligrosas, este proyecto de titulación se centrará solamente en la ciudad de Guayaquil y su incremento delincencial que se debe mayormente por las peleas entre grupos de narcotraficantes los cuales mandan a matar a través del sicariato a miembros de bandas contrincantes, mayormente personas con registros penales dando como resultado muertes de personas que fueron interceptadas por una bala perdida la estar pasando por la calle en el momento del atentado.

En aproximadamente los últimos 5 años la ciudad de Guayaquil ha sufrido un aumento considerable en la delincuencia que atemoriza día a día a la población guayaquileña ya sea en los barrios urbanos o hasta en los semáforos que aun uno estando en un vehículo, no hay seguridad que en cualquier momento las personas sean asaltadas. Pero la verdad es que no en cada zona residencial de Guayaquil existe el riesgo de ser robado, donde más se puede escuchar este tipo de acontecimientos pasa en sectores los cuales los residentes no constan con un estatus social medio-alto, sino que estos constan de bajo a medio-bajo.

La razón por la cual en estos sectores de la ciudad pasan más este tipo de situaciones es que al contener alquileres o terrenos baratos, son accesible para que personas pertenecientes a bandas narcoterroristas puedan realizar sus operaciones sin que sean tan controlados por la policía. Pero no todo lugar

que contenga alquileres y terrenos bajos en costos significa que la delincuencia predomina en ese sector, por lo cual el modelo predictivo de este proyecto se enfocara en predecir qué lugares podrían volverse peligrosos y a partir de esta información tomar una decisión si mudarse o no en este sector.

El presente trabajo de titulación está constituido principalmente por cuatro capítulos, CAPITULO I el cual va a tratar con relación a la problemática de la cual surgió el tema de investigación junto a su objetivo general y objetivos específicos. El CAPITULO II consta del marco teórico en donde se encontrará los conceptos necesarios para el entendimiento de lo que tratará el proyecto. El CAPITULO III será sobre la metodología de la investigación que tratara de las metodologías que se investigaron para obtener los datos necesarios y poder después usarlos. Por último, se tendrá el CAPITULO IV el cual habla de la metodología de la investigación donde se detallará los procesos y como se llevó a cabo el proyecto junto a sus algoritmos y métodos.

# **CAPÍTULO I**

## **EL PROBLEMA**

### **PLANTEAMIENTO DEL PROBLEMA**

#### **Ubicación del Problema en un Contexto**

Actualmente en el Ecuador la delincuencia ha tenido un auge mayormente proveniente por el hecho que las personas que comenten estos actos violentos, en su mayoría hombres jóvenes es debido a las altas tasas de desempleo juvenil, la impunidad en el sistema judicial y el acceso fácil al alcohol, las drogas y las armas de fuego, ahora a esto se le aumenta el hecho de que Ecuador es uno de los países de Latinoamérica que es la red de narcotráfico más grande y extensa del mundo y esto influencia a los jóvenes a involucrarse en dicho negocio para poder ganar algo de dinero realizando actividades ilegales poniendo en peligro a personas las cuales no tienen relación alguna con las drogas y terminan siendo víctimas de peleas entre bandos de grupos de carteles de droga. (Mayra Buvinic, 2005)

#### **Causas y Consecuencias del Problema**

Las causas que existen por la cual la ciudad de Guayaquil se haya vuelto tan peligrosa son varias, una es el hecho que Guayaquil al ser la principal ciudad portuaria del Ecuador (España, 2021) muestra varias oportunidades para los narcos a poder transportar ya sea a Europa o Estados Unidos toneladas de droga en contenedores lo cual representa una pelea de poder por el sector portuario poniendo en riesgo a las personas que viven cerca de los puertos marítimos ubicados en la ciudad de Guayaquil. Otra causa por la cual la gente se siente insegura en la ciudad se deriva del hecho

que, al intentar transportar la droga por los contenedores, las fuerzas armadas y militares incautan las toneladas de cocaína y proceden a quemarlas, dañando así el negocio de exportación de las drogas hacia países extranjeros. (PlanV, 2018) Otra causa por el aumento de la delincuencia se debe a que en Guayaquil al ser una de las ciudades más grandes del país cuenta con varios sectores mayormente con residentes de un estado social económico bajo o medio-bajo donde las personas son tentadas a cometer robo, hurto o en casos más extremos asesinato, esto tras considerar la ganancia del costo-beneficio de lo que conseguiría al cometer el acto delictivo, como por ejemplo el robo de una billetera o bolso esto se debe a la diferencia que el agresor y el agredido tienen como ingresos monetarios lo cual tenta al agresor a cometer actos violentos. (Mayra Buvinic, 2005)

Las consecuencias que son originadas por las causas mencionadas son la agresión de los narcos hacia las fuerzas del orden como secuestros, asesinatos y hasta destrucción de Unidades de Policía Comunitaria (UPC) con carros bomba como lo fue el primero de noviembre del 2022 en la cual 5 policías fueron asesinados teniendo que poner al país en estado de excepción y toque de queda por temor que vuelvan a haber más actos violentos. (Mundo, 2022)

Una consecuencia que se originó debido a la agresión a los policías es imponer un orden más estricto en la penitenciaría del litoral el cual alberga una cantidad de 8500 de privados de libertad, siendo el epicentro de la confrontación entre bandas delictivas por el negocio del narcotráfico y el liderazgo. Se estima que se han creado al menos 20 bandas delictivas dentro de las celdas con un número de integrantes de 40000 entre locales y

extranjeros, por lo cual se ha decidido movilizar a los reos a otros centros de reclusión para separar y reubicar los integrantes de estas bandas e intentar traer la paz de nuevo al país. (Román, 2022)

Esta confrontación entre fuerzas de la justicia y bandas de narcos lo que pueden traer a futuro como consecuencia, es más personas alcanzadas por el fuego entrecruzado por parte de la violencia que los narcos realizarán hacia las fuerzas policiales y militares en forma de venganza así nunca llegando a la tranquilidad en la ciudad y el país.

### **Delimitación del Problema**

El modelo predictivo del presente trabajo de titulación será generado con los datos obtenidos del estudio hecho en las zonas residenciales de la ciudad de Guayaquil que representen viviendas accesibles para familias de un estado socio-económico bajo y medio-bajo.

|         |   |
|---------|---|
| Campo   | Procesamiento de datos en las zonas con un alto índice de peligrosidad.   |
| Área    | Minería de datos  |
| Aspecto | Un modelo predictivo que indique las zonas más peligrosas de Guayaquil.   |
| Tema    | Modelo predictivo que analice cuales son las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y a partir de esta información predecir qué sectores en desarrollo tendrán el mismo nivel de delincuencia |



## **Formulación del Problema**

En la actualidad del país ecuatoriano, uno de los temas más comunes entre su población es como el país se ha vuelto una patria tan insegura con el paso de los últimos años, lo cual ha dado consecuencias como las muertes de muchas personas inocentes, ajenas a problemas entre peleas de bandas narcotraficantes lo cual ha resultado un país en el que se vive con la desconfianza que en cualquier momento va a ser asaltado o hasta matado en cualquier zona, incluyendo hasta las zonas residenciales que la década pasada se consideraban medianamente seguras o hasta completamente seguras.

Por lo cual para la formulación del problema se procede a contestar las siguientes preguntas, ¿Cómo podrá mejorar la toma de decisiones al momento de querer comprar o mudar una casa? ¿Cuáles son las herramientas que serán necesarias para la creación del modelo predictivo? ¿Cómo podría la minería de datos predecir qué zonas son consideradas peligrosas?

## **Evaluación del Problema**

Los seis aspectos generales de evaluación de mi problema son los siguientes:

**Delimitado:** El problema de investigación es delimitado ya que se ha especificado en la descripción del problema como el incremento de la delincuencia ha afectado a la población guayaquileña, en el tiempo de los últimos años en comparación de hace una década y en el espacio que se refiere a las zonas de bajo y medio-bajo índice de estatus socio-económico en la ciudad de Guayaquil.

**Evidente:** El aspecto de evidente también es claro en el planteamiento del problema de investigación ya que a simple vista se puede ver como se manifiesta el temor de las personas con tan solo observar en cómo salir fuera de sus hogares a altas horas de la tarde o noche les puede llegar a provocar, junto a la cantidad de noticias que se comparten diariamente sobre como personas de todas las edades y genero resultan heridas o hasta muertas por culpa de la delincuencia.

**Claro:** El planteamiento del problema forma parte del aspecto de claro porque tal como indica su nombre, el problema describe con claridad la situación en que la ciudad de Guayaquil se encuentra frente a la delincuencia que sufre diariamente las personas junto a sus razones y consecuencias que se generarían en las personas si la violencia en el país sigue.

**Relevante:** El problema de investigación de igual manera contiene el aspecto de relevante, al explicar cómo trae conflicto a la vida de los guayaquileños que antes podían gozar de una salida en la tarde o noche sin tener tanto temor en poder terminar robado o muerto. También se explica la importancia que este problema trae consigo y el por qué es relevante que se realice este problema de investigación.

**Original:** El problema de investigación muestra originalidad al tratarse de algo tan común pero no tan comúnmente investigado científicamente, más las personas se enfocan en el problema y lo que se escucha en las redes sociales y noticieros por lo cual el enfoque de esta investigación es la predicción de zonas en crecimiento y ver si son seguras para vivir o no en el futuro.

**Factible:** Forma parte del aspecto factible al ser un problema que puede tener una solución a futuro, esta puede ser conseguida a través de movimientos

sociales, decisiones políticas o hasta con la ayuda de una herramienta tecnológica la cual ayude a los guayaquileños a decidir una vivienda que sea lo suficientemente seguro para sus necesidades.

## **OBJETIVOS**

### **OBJETIVO GENERAL**

Diseñar un modelo predictivo que muestre las áreas propensas a convertirse en barrios peligrosos en comparación con las actualmente más peligrosas de la ciudad de Guayaquil.

### **OBJETIVOS ESPECÍFICOS**

- Recopilar y analizar las variables que permiten calificar a un sector o barrio peligroso para el diseño de un modelo predictivo.
- Elaborar un modelo predictivo utilizando minería de datos para calificar a los sectores que pueden volverse peligrosos.
- Valorar la aproximación de las predicciones que se generaron a través del modelo predictivo.

### **ALCANCES DEL PROBLEMA**

El problema de la investigación se lo abordará con herramientas de minería de datos los cuales permiten probar varias corridas de datos que serán estudiados y analizados para finalmente escoger el resultado que más se ajuste al tema de estudio de esta investigación científica demostrando el uso de la metodología escogida de minería de datos y así poder decidir si el tema de investigación logro o no cumplir todos los objetivos planteados.

La información que se usara para este trabajo investigativo va a ser obtenida de la página web de la Ecu911, la cual cuenta con una sección en

donde se muestra las cifras que se tiene registrada de las emergencias hechas por las personas en un periodo editable de tiempo, donde también se puede filtrar la información que se necesita a través de filtros como provincia, cantón, zona, tipo de servicio, tipo de subservicio, entre muchas otras opciones más.

## **JUSTIFICACIÓN E IMPORTANCIA**

La minería de datos es una metodología la cual se usa muchas veces para la toma de decisiones ya que este método genera varias corridas o patrones los cuales muestran diferentes comportamientos con el objetivo de poder predecir lo más exactamente posible.

El uso de la minería de datos en este trabajo de titulación es fundamental porque ayuda a proveer las herramientas necesarias para la toma de decisión y predicción de si una zona es habitable o no dependiendo del nivel de delincuencia en sectores más poblados. Esto ayuda a las personas que estén buscando una casa o departamento el cual vivir que no presente tanta delincuencia ni violencia para tener una vida más tranquila.

Este proyecto de titulación es relevante socialmente ya que aporta una ayuda a la sociedad a través de un modelo predictivo al brindar una herramienta o forma que facilite la toma de decisiones del poblado de Guayaquil al momento de decidir qué sector quiere residir.

## **HIPÓTESIS O PREGUNTA DE INVESTIGACIÓN**

El presente trabajo de titulación plantea la siguiente hipótesis:

Un modelo predictivo el cual utiliza la minería de datos para predecir los sectores en crecimiento que tendrán un nivel alto de delincuencia.

## **VARIABLES DE LA INVESTIGACIÓN**

**Variables independientes:** Modelo predictivo

**Variables dependientes:** Zonas con mayor índice de peligrosidad y sectores en desarrollo que tendrán el mismo nivel de delincuencia

## **CAPÍTULO II**

### **MARCO TEÓRICO**

Para entender de mejor manera como se usa la minería de datos y sus metodologías, en este capítulo se explicará las diferentes metodologías que existen acerca de la minería de datos junto a sus diferencias entre cada técnica, las definiciones conceptuales, normas establecidas, estándares, leyes, reglamentos y criterios para un entendimiento óptimo acerca del uso de minería de datos para el desarrollo de un modelo predictivo.

#### **MODELOS PREDICTIVOS**

El modelado predictivo o modelos predictivos es una técnica que usa las estadísticas y resultados conocidos para procesar y crear modelos que se pueden usar para predecir resultados futuros, dentro de lo razonable. A medida que el modelado predictivo es más rentable y está disponible fácilmente, los clientes exigen experiencias avanzadas que les lleven a mejores decisiones y acciones. (BrianBlanchard, 2022)

Según la unir (La Universidad en Internet) los análisis predictivos o modelos predictivos se basa en estimar eventos futuros en función de datos históricos, a los que se les aplican diferentes técnicas analíticas, estadísticas y de aprendizaje autónomo. Son modelos matemáticos que predicen el comportamiento de una variable en función de un conjunto de otras variables. Cuanto más relacionadas estén el conjunto de variables predictoras con la variable a predecir (correlación), más exactas serán las predicciones. (Unir, 2020)

Para la construcción de estos modelos, una vez desarrollado el algoritmo predictivo, es necesario disponer de un conjunto de datos históricos. Para ello se dividen estos en dos conjuntos: uno de datos de entrenamiento y otro de prueba. Al modelo se le pasan como entrada los datos de entrenamiento para calibrar la predicción y, posteriormente, los datos de prueba. Después se compara el resultado de la predicción con los valores reales (históricos) para comprobar su exactitud. (Unir, 2020)

## **MINERÍA DE DATOS**

La minería de datos, también conocida como minado de datos o en inglés data mining según Elena Bello son un conjunto técnicas y tecnologías las cuales ayudan a estudiar y explorar grandes volúmenes de datos de manera automática y semiautomática con el fin de encontrar comportamientos mediante patrones de los datos estudiados. La minería de datos se desarrolló con la intención de ayudar a comprender grandes cantidades de datos y que puedan ser utilizados para extraer conclusiones para contribuir en la mejora de tomas de decisiones. (Bello, 2021)

Otra definición de la minería de datos proviene de la documentación que Microsoft provee que está redactada en colaboración de cuatro usuarios, estos autores explican que la minería de datos detalla el proceso de detectar la información procesable de los conjuntos grandes de datos. Estos patrones no detectan la exploración tradicional de los datos ya que las relaciones son demasiado complejas o porque hay demasiados datos. (Minewiskan, 2022)

Una última definición de la minería de datos propuesta por la Revista Iberoamericana para la Investigación y el Desarrollo Educativo. A través de diversas técnicas, se extrae información de una base de datos para generar conocimiento, el cual puede ser expresado a través de conceptos, reglas, leyes, patrones, entre otros. (Domínguez y otros, 2022)

La minería de datos no es un término que recién hace aparición, es un término que ha comenzado a sonar más con la llegada de la transformación digital haciendo dar cuenta lo necesario y poderoso que puede ser el valor de los datos. Estos datos se acumulan y registran diariamente en diferentes instituciones como negocios y empresas en donde ha pasado a ser una materia prima que se transforma en conocimiento valioso para estrategias y alcanzar objetivos en el ámbito de los datos. (inesdi, inesdi, 2021)

Muchas personas confunden la minería de datos con la estadística, a primera vista si pareciera que son sinónimos mostrando cierto paralelismo, pero la verdad es que la minería de datos usa un gran volumen de datos con el objetivo de extraer conocimiento. Por ejemplo, la minería de datos ayuda a concluir ciertos resultados que se consiguen por medio de técnicas estadísticas las cuales ayudan a establecer un conjunto de reglas en antecedentes y de esta manera llegar a una predicción aproximada. La minería de datos tiene la gran diferencia con la estadística que no arroja datos, sino conocimiento nuevo a partir de datos históricos originado por una herramienta o método automático. (inesdi, inesdi, 2021)

El proceso de la minería de datos tiene diferentes estados o etapas por las cuales tiene que pasar para poder obtener nuevo conocimiento originado



de los datos estadísticos y datos históricos obtenidos previamente. Los estados son los siguientes: (inesdi, inesdi, 2021)

- Selección y preparación de datos: se determinan las fuentes, se transforma y estandarizan los datos, se corrigen y se almacenan. (IBM, IBM, 2021)
- Creación del modelo de minería de datos: Un modelo de minería de datos se crea a partir de un conjunto específico de datos de entrada, se debe especificar donde residen los datos de entrada, que campos de los datos son apropiados, que valores se deben utilizar para la función de minería determinada que se está utilizando y donde almacenar el modelo final. (IBM, IBM, 2021)
- Evaluación del modelo y análisis de su calidad: Se puede probar un modelo con la técnica elegida como regresión o clasificación y analizar la calidad del modelo. (IBM, IBM, 2021)
- Difusión: Se visualiza los resultados para analizarlos e interpretarlos. (IBM, IBM, 2021)

## **TÉCNICAS DE MINERÍA DE DATOS**

Para aplicar modelados de minería de datos existen distintas técnicas las cuales son usadas diariamente por empresas. (inesdi, inesdi, 2021)

### **Árbol de decisión**

Tiene este nombre porque tiene una estructura parecida a un árbol en donde se utilizan nodos de dos tipos:

- Los puntos de decisión

- Los puntos de azar

Los problemas y la secuencia de los árboles de decisión se plasman en estos árboles, donde un nodo es un punto de unión conectado por ramas. El árbol se crea de izquierda a derecha, pero se evalúa de forma inversa, simplemente porque a la izquierda se encuentra la decisión y a la derecha los resultados. (inesdi, inesdi, 2021)

El árbol de decisión consta de 4 elementos:

- Puntos de decisión: se representan con un cuadrado. Aquí el decisor elige una alternativa de acción entre un número finito de ellas que son representadas por las ramas cuyos costes asociados se escriben sobre ellas. Las ramas escogidas pueden acabar en otro punto de decisión, en uno de azar o en un resultado. (inesdi, inesdi, 2021)
- Puntos de azar: se dibujan con un círculo e indican que un suceso aleatorio se espera en este punto del proceso. Desde aquí también surgen ramas. (inesdi, inesdi, 2021)
- Ramas: en el argot del big data se definen como alternativas cuando salen de los puntos de decisión y como estados de la naturaleza cuando salen de los puntos de azar. En este último caso, se les asigna unas probabilidades determinadas. (inesdi, inesdi, 2021)
- Resultado: al final tenemos que decidir qué decisión tomar en función del resultado obtenido proveniente de cada rama. (inesdi, inesdi, 2021)

## **Red neuronal**

Esta técnica de data mining se basa en el funcionamiento de las neuronas del ser humano, pues el cerebro tiene millones que se conectan

entre sí en un proceso llamado “sinapsis”. Esta red neuronal artificial se parece tanto a una biológica que cuenta con nodos de entrada (reciben información del exterior), nodos de salida (transmiten información al exterior) y nodos ocultos (intercambian información con otros nodos de la red). (inesdi, inesdi, 2021)

Cuando estos nodos están definidos se pasa a la fase de aprendizaje donde se asignan diferentes valores a estos nodos hasta encontrar respuestas, pues es la propia red la que los crea, modifica o elimina automáticamente. La principal ventaja de esta técnica de data mining es su capacidad para trabajar con datos incompletos. (inesdi, inesdi, 2021)

### **Modelado estadístico**

Se basa en las relaciones entre variables en los datos mediante ecuaciones matemáticas para predecir resultados. Es la más antigua de las técnicas de minería de datos, ya que se empezó a desarrollar en el siglo XVII con métodos más arcaicos, pero la esencia era la misma que en la actualidad. Si es tan antigua, es porque es una rama de las matemáticas que se fue introduciendo al mundo de los datos a medida que fueron incorporándose en la sociedad actual. (inesdi, inesdi, 2021)

### **Reglas de asociación**

Esta técnica permite encontrar las combinaciones de artículos que ocurren con mayor frecuencia en una base de datos y la importancia de las mismas. (inesdi, inesdi, 2021)

Un ejemplo de esta técnica de data mining es el cliente que va a comprar un artículo y su intención de compra se asocia con la de otros consumidores en la base de datos, o incluso se le muestran otros productos basándose en su historial. Los datos se agrupan en forma de lista, en una representación vertical o en una horizontal. (inesdi, inesdi, 2021)

### **Agrupamiento (Clustering)**

Se agrupan elementos en un conjunto de datos que, a su vez, están agrupados en subconjuntos distintos. El objetivo es que los elementos de una misma clase tengan grandes similitudes entre sí, mientras que los que pertenezcan a una clase distinta cuenten con el menor parecido posible. (inesdi, inesdi, 2021)

Hay muchos tipos de clustering, pero los más frecuentes son dos:

- Clustering jerárquico: un objeto está más relacionado con los objetos que tiene cerca que con los objetos lejanos. (inesdi, inesdi, 2021)
- Clustering basado en la densidad: se agrupan los objetos en clústeres siempre y cuando los elementos más cercanos estén dentro de un umbral establecido. (inesdi, inesdi, 2021)

### **Algoritmo genético**

Al igual que la red neuronal está basada en las neuronas humanas, el algoritmo genético está basado en la teoría de la evolución. En esta técnica de data mining se intenta replicar el comportamiento biológico de la selección natural y la genética. El algoritmo cuenta con una población inicial de datos que representan ciertos resultados (cromosomas) y que contienen bits (genes). Estos pasan juntos a la fase de evaluación donde se le asignará un

porcentaje en función de la aptitud. Los más aptos siguen y los demás no, igual que en la teoría de Charles Darwin. Después de esto, los datos se cruzan o mutan y el proceso se repite hasta que se llega al resultado esperado o hasta que se para manualmente. (inesdi, inesdi, 2021)

### **Regresión lineal**

La regresión lineal es otra de las técnicas de minería de datos más utilizadas en un sector que no para de crecer debido a la transformación digital. En ella, se relacionan dos variables continuas, concretamente, las variables de predicción y de respuesta. (inesdi, inesdi, 2021)

Se habla de regresión lineal cuando existe solo una variable de predicción y de regresión múltiple cuando hay más de una. Sea lineal o múltiple, es una variable independiente mientras que la de respuesta depende de la anterior. (inesdi, inesdi, 2021)

### **Redes bayesianas**

Representan ciertas incertidumbres que están asociadas a nodos que reproducen variables aleatorias, las cuales se asocian a su vez a un condicionante externo. Para esto, existen los llamados “clasificadores bayesianos”, que organizan cada variable y consiguen plasmar los condicionantes de tal manera que sean muy sencillos de leer. Son muy característicos en la medicina para diagnósticos graves. Se utilizan las redes bayesianas para descartar enfermedades rápidamente. (inesdi, inesdi, 2021)

## **METODOLOGÍAS DE MINERÍA DE DATOS**

La minería de datos es parte de macro datos o también conocido como Big data, que gestiona y analiza conjuntos extensos de datos. La minería de datos tiene varios usos en diferentes campos de estudio. Los algoritmos del Data Mining pueden realizar tres tipos de análisis: descriptivos, predictivos o prescriptivos. (Veigler, 2021)

- **Analítica descriptiva.** Este tipo de análisis de algoritmos descriptivos buscan nexos con el fin de describir datos. (Veigler, 2021)
- **Analítica predictiva.** El análisis predictivo, permite predecir cómo evolucionará y se comportará o cuáles serán las próximas tendencias futuras dependiendo de los datos históricos. (Veigler, 2021)
- **Analítica prescriptiva.** En este análisis se usan decisiones obtenidas de predicciones a partir de datos históricos. Estas decisiones se escogen de forma automática, porque combinan la optimización matemática con los servicios de gestión empresarial. (Veigler, 2021)

### **KDD (Knowledge Discovery in Databases)**

La metodología KDD o el Descubrimiento de Conocimiento de Base de datos en español se refiere al uso de la minería de datos que extrae conocimiento al usar una base de datos en conjunto con cualquier pre procesamiento requerido en la base de datos. (Zambrano y otros, 2022)

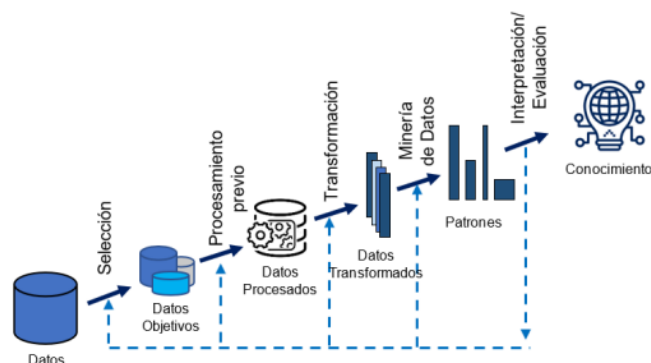
Esta metodología consta de cinco pasos principales:

- **Selección de datos:** Consiste en crear un conjunto de datos o muestras de datos. (Zambrano y otros, 2022)

- Preprocesamiento y limpieza: Incluye limpieza y eliminación de ruido, campos vacíos o elementos repetidos. (Zambrano y otros, 2022)
- Transformación: En esta etapa la metodología transforma los datos que se tienen al usar métodos que reducen su volumen. (Zambrano y otros, 2022)
- Minería de datos: En esta etapa se procede a usar algoritmos pertenecientes a aprendizaje de máquina para encontrar patrones que dependen del objetivo. (Zambrano y otros, 2022)
- Interpretación: Consiste en la evaluación e interpretación de los resultados obtenidos. (Zambrano y otros, 2022)

**Figura 1**

*Grafico describiendo los pasos de la metodología KDD*



*Nota: Grafico obtenido de (Zambrano y otros, 2022)*

### **SEMMA (Sample Explore Modify Model Assess)**

La metodología SEMMA como su nombre lo indica el acrónimo para las cinco fases por las que pasa esta metodología los cuales son muestra, explorar, modificar, modelar y evaluar. Esta metodología esta propuesta por SAS Institute Inc y se la defina como el proceso de selección, la exploración y

el modelamiento de cantidades enormes de datos para obtener patrones de negocios o situaciones desconocidas. (León, s.f.)

- Sample: Entrada de datos, ejemplos, partición de datos. (León, s.f.)
- Explore: Exploración distribuida, múltiples particiones, intuición, asociación, selección de variables. (León, s.f.)
- Modify: Transformación de variables, filtros a los datos fuera de rango, agrupamiento, ruido. (León, s.f.)
- Model: Regresión, árboles, redes neuronales, etc. (León, s.f.)
- Assess: Evaluación, medidas, reportes. (León, s.f.)

## **CRISP-DM**

CRISP-DM compuesta de sus siglas en inglés significa Cross-Industry Standard Process for Data Mining es un modelo de proceso de minería de datos que en las empresas privadas y públicas, la metodología CRISP-DM es de las más usadas por varias empresas y consta de seis pasos, que son los siguientes: (Galán, 2015)

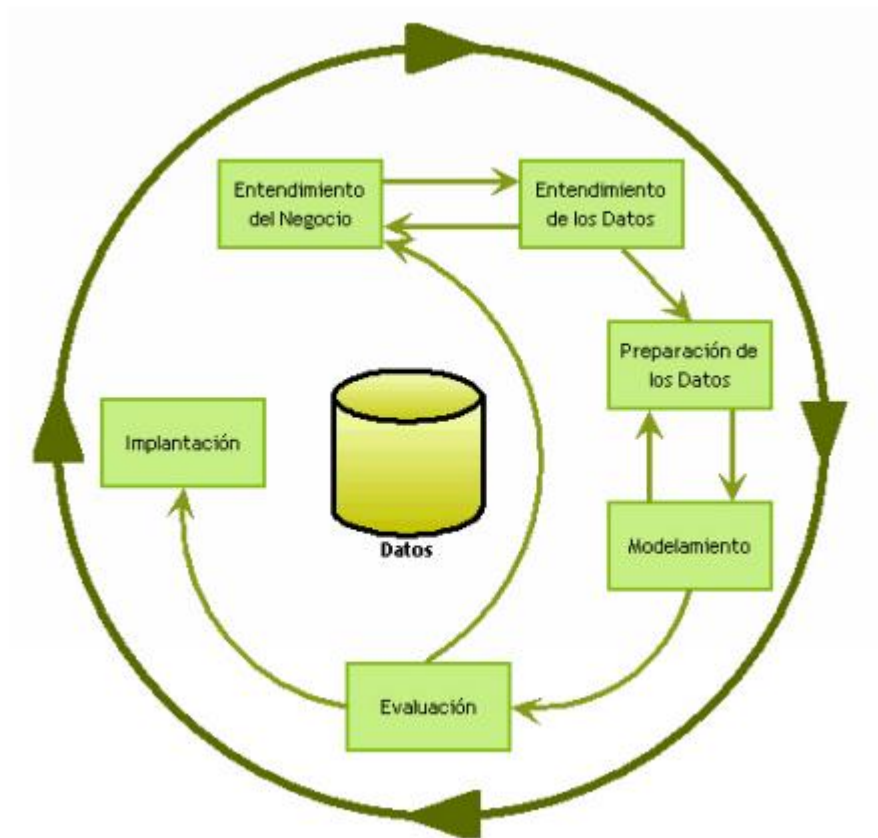
- Comprensión del negocio: Entendimiento de los objetivos y requerimientos del proyecto y definición del problema de minería de datos. (León, s.f.)
- Comprensión de los datos: Obtención del conjunto inicial de datos, exploración del conjunto de datos, identificar las características de calidad de los datos e identificar los resultados iniciales obvios. (León, s.f.)
- Preparación de datos: Selección de datos y limpieza de datos. (León, s.f.)



- Modelamiento de datos: Implementación en herramientas de minería de datos. (León, s.f.)
- Evaluación: Determinar si los resultados coinciden con los objetivos del negocio e identificar los temas de negocio que deberían abordarse. (León, s.f.)
- Despliegue: Instalar modelos resultantes en la práctica y configurar la minería de datos de forma repetida o continua. (León, s.f.)

**Figura 2**

*Gráfico de la metodología de CRISP-DM*



*Gráfico que explica la metodología de CRISP-DM, obtenido de (Galán V. , 2015)*

## COMPARACION ENTRE KDD, SEMMA Y CRISP-DM

**Tabla 1**

Comparación de métodos de minería de datos

| KDD                                      | SNEMMA       | CRISP-DM                  |
|--|--------------|---------------------------|
| xxx                                      | xxx          | Conocimiento del negocio  |
| Selección de datos                       | Muestra      | Conocimiento de los datos |
| Preprocesamiento y análisis de los datos | Exploración  |                           |
| Transformación                           | Modificación |                           |
| Minería de datos                         | Modelo       |                           |
| Interpretación de resultados             | Evaluación   |                           |

*Nota: Comparación de metodologías de minería de datos (León, s.f.)*

## LENGUAJES DE PROGRAMACION PARA MINERIA DE DATOS

### Python

Python tiene como característica ser un lenguaje de programación flexible con aspectos de multiplataforma y multiparadigma que tiende a sobresalir por su código fácil de entender y limpio. La licencia de código abierto permite su utilización en distintos contextos sin la necesidad de abonar por ello y se emplea en plataformas de alto tráfico como Google, YouTube o Facebook. (Universia, 2020)

## Para que sirve Python

El objetivo del lenguaje Python es la automatización de procesos para ahorrar tiempo al reducir procesos a través de pocas líneas de código. Este lenguaje de programación puede usarse en los siguientes campos:

- **Ciencia de datos:** El poder de las bibliotecas Python desarrolladas para el análisis y visualización de datos es asombroso. Con una biblioteca de visualización de datos de Python, puede crearse una amplia variedad de gráficos y representaciones visuales de todo tipo. (Universia, 2020)
- **Aprendizaje automático.** Python es una herramienta esencial para todos los desarrolladores que quieran sumergirse en el campo del machine learning. Una de las bibliotecas más populares que utilizan los desarrolladores de todo el mundo para trabajar con Python aplicado al aprendizaje automático es TensorFlow. Se trata de un centro de recursos gratuito de código abierto desarrollado por el equipo de Google Brain. Esta biblioteca se utiliza para investigación y producción en Google. (Universia, 2020)
- **Desarrollo web.** Python se utiliza en el campo del desarrollo web para construir el back-end de aplicaciones web. (Universia, 2020)
- **Educación en Ciencias de la Computación.** Python se usa ampliamente como herramienta de enseñanza porque es fácil de aprender: su sintaxis es simple y se puede aprender rápidamente. Es potente y permite a los estudiantes comenzar a adquirir habilidades valiosas para sus carreras de inmediato, y es versátil, ya que admite varios paradigmas de programación como la programación imperativa,

la programación funcional, la programación procedimental y la programación orientada a objetos.

- **Visión por ordenador y procesamiento de imágenes.** Permite a los desarrolladores integrar funciones de detección de visión dentro de las aplicaciones de manera sencilla.
- **Desarrollo de juegos.** Los juegos crean recuerdos atemporales y seguirán formando parte de nuestra sociedad en los próximos años. Python respalda la innovación aplicada a la creación de juegos.
- **Medicina y Farmacología.** Python también tiene aplicaciones asombrosas en el campo médico que mejoran la capacidad de brindar diagnósticos y tratamientos precisos y eficientes a los pacientes.
- **Biología y Bioinformática.** Sus aplicaciones en estos campos tienen que ver con el procesamiento de secuencias de ADN, la simulación de dinámica y genética de poblaciones y el modelado de estructuras bioquímicas.
- **Neurociencia y Psicología.** Tal y como se publica en un reciente artículo, "la computación se está volviendo esencial en todas las ciencias, para la adquisición y análisis de datos, la automatización y la prueba de hipótesis a través del modelado y simulación".
- **Astronomía.** Python también tiene aplicaciones en Astronomía y Astrofísica. Sus principales aportaciones a estas áreas son AstroPy, SunPy y SpacePy.
- Otras áreas como robótica, vehículos autónomos, negocios, meteorología y desarrollo de interfaces gráficas de usuario también se benefician del uso de Python.

## **Características de Python**

Atrapa a los usuarios por su sencillez, legibilidad y exactitud en la sintaxis, entendida como el conjunto de reglas que se siguen al escribir un código pues se trata de lenguaje como cualquier otro empleado en la comunicación de ideas, aunque a nivel informático. Con unas pocas líneas de código podrás programar algoritmos complejos que arrojen resultados sofisticados, volviéndolo un lenguaje práctico y ahorrativo en materia de tiempo. Además, posee dialectos -variantes que se adaptan a otros lenguajes- como Python que se utiliza para escribir en Java, el lenguaje de programación más empleado en sitios y aplicaciones por su servicio completo.

Para los fanáticos de la estética informática, podría catalogarse como un lenguaje elegante y casi minimalista. Comprueba los errores sobre la marcha para solucionarlos cuando afectan a la memoria, manteniendo la integridad de tu matriz y evitando las complicaciones a la hora de escribir el código. Con casi 30 años en el mercado, su antigüedad lo volvió un lenguaje sólido que maduró entre la comunidad de adeptos, siendo un indicador clave sobre su calidad y practicidad de uso. Finalmente, su enfoque en la funcionalidad es la causa principal que nos impulsa a recomendarte el aprendizaje de Python si estás pensando en ingresar al mundo de la programación.

## **CAPÍTULO III**

# **METODOLOGÍA DE LA INVESTIGACIÓN**

### **TIPO DE INVESTIGACIÓN**

Existen diferentes tipos de metodologías de la investigación, como la cuantitativa, cualitativa y hasta la mixta. Cada metodología tiene su propia técnica, diseños, métodos e instrumentos que van de acuerdo con la naturaleza del objeto de estudio. (Monje Alvarez, 2011)

Los métodos de investigación están conformados por procedimientos específicos para recopilar y analizar datos, por eso para realizar un estudio se tiene que decidir que técnica de investigación utilizar, la cuantitativa o cualitativa, aunque esta decisión dependerá más del objeto y ámbito de estudio, los tipos de datos que se vayan a utilizar y las personas o elementos de donde se obtendrán los datos. (Santander, 2021)

Para poder diferenciar la metodología cuantitativa de la cualitativa, a simple vista puede ser diferenciada por la representación de los datos recogidos como palabras o números, si se usan los números la investigación será posiblemente cuantitativa en caso de que se usen más palabras para describir los datos, el tipo de investigación sería cualitativa. (Santander, 2021)

La metodología de investigación cuantitativa es usada para comprensión de frecuencias, patrones, promedios, probar o confirmar teorías, entre muchas otros campos y funciones más. (Santander, 2021)

De las herramientas más empleadas para la recopilación de dato son:

- Encuestas o cuestionarios: las encuestas son preguntas que tienen como respuesta una cerrada o de escoger opciones al presentar preguntas cerradas que no generen una conversación. (Santander, 2021)
- Experimentos: En los experimentos se incluyen pruebas que confirmen las hipótesis planteadas por el autor usando de relaciones de causa y efecto. (Santander, 2021)
- Observación: La herramienta de observación trata de contar y generar reportes del número de ocasiones un fenómeno o evento en concreto pasa. (Santander, 2021)

## **METODOLOGIA DE MINERIA DE DATOS**

En el Capítulo II el cual habla del marco teórico de la investigación de titulación, se explica las diferentes metodologías investigadas las cuales son usadas específicamente para lo que es la minería de datos y para la presente investigación se ha decidido usar la metodología KDD por sus siglas Knowledge Discovery in Databases. Este método consta de 5 etapas las cuales serán explicadas de manera sencilla y superficial en el presente capítulo, mientras en el capítulo siguiente se procederá a tener una explicación más detallada de las etapas de KDD.

### **Selección de datos**

En esta etapa como indica su nombre se selecciona de una fuente confiable o se crea un conjunto de datos objetivo, sobre el cual se realizará los pasos a seguir de la metodología KDD. Para esta etapa la fuente de datos

que se va a usar es obtenida de un portal usando PowerBI de la Ecu911 en donde se especifica detalladamente todos los tipos de emergencia en que los Guayaquileños han acudido al Ecu911. Estas emergencias se pueden filtrar por provincia, ciudad, cantón y hasta parroquias junto con un rango de fechas que uno desee. La información obtenida del portal mencionado puede ser descargado en formato de hoja de cálculo lista para ser usado en las siguientes etapas de la metodología KDD.

### **Preprocesamiento y limpieza**

En la etapa de preprocesamiento y limpieza es el paso en donde se analiza la calidad de los datos obtenidos eliminando el ruido de los datos, datos repetidos o datos nulos que normalmente aparecen al obtener el conjunto de datos de personas lo cual es normal al ser un error humano que no se puede controlar en grandes volúmenes de información, pero en este caso en específico del presente trabajo, la limpieza de los datos no aplicaría a la eliminación del ruido o datos repetidos ya que la información es proporcionada por una institución la cual revisa sus datos antes de publicarlos al público general.

De igual manera, aunque los datos ya se presenten limpios, se procede a realizar métodos de limpieza de datos y de preprocesamiento verificando que no se elimina ningún valor del set de datos al ser datos ya revisados y limpios.

### **Transformación**

La transformación de los datos explica como los datos al ya estar limpios sin ningún valor repetido o no valido, se procede a reducir el volumen



de datos al eliminar variables y atributos que son insignificantes o no contribuyen con informacion con respecto al problema tratado.

En este trabajo de titulación la reducción y transformación de datos no útiles pueden ser filtrados a través del portal de la Ecu911 al poder suprimir de emergencias ciudadanas las cuales no tienen relación con la inseguridad o delincuencia presente en la actualidad, como por ejemplo los temblores o inundaciones.

### **Minería de datos**

En la etapa de minería de datos se procede a utilizar los algoritmos de aprendizaje supervisado escogidos que pertenezcan a algoritmos de regresión como regresión lineal, random forest, máquina de vectores, regresión lógica, entre muchas otras más. Tras realizar diferentes corridas con los algoritmos escogidos se obtiene predicciones las cuales van a ser analizadas posteriormente.

### **Interpretación**

Con las predicciones obtenidas en el paso anterior, lo siguiente es la comparación, interpretación y análisis de los resultados para determinar la aproximación de las predicciones obtenidas y así poder concluir los objetivos planteados.

## **POBLACIÓN Y MUESTRA**

### **Población:**

La población total que se consideró para este proyecto de titulación consta de los habitantes que conforman a la ciudad de Guayaquil, más

específicamente la población que forma parte de las parroquias urbanas que conforman la ciudad de Guayaquil.

La ciudad de Guayaquil está conformada por 16 parroquias urbanas las cuales según la INEC en el año 2010 en donde se realizó el último censo global oficial del Ecuador, Guayaquil tenía una población de 2'350.915 por lo cual esta será la población que se usará en este proyecto de titulación al ser el último registro brindado por parte de la INEC. (INEC, 2010)

## EL TAMAÑO DE LA MUESTRA

$$n = \frac{m}{e^2 (m - 1) + 1}$$

$$n = \frac{2350915}{(0.06)^2 (2350915 - 1) + 1}$$
$$n = \frac{2350915}{(0.0036)(2350914) + 1}$$
$$n = 278$$

El tamaño de la muestra tras usar la fórmula de cálculo de la muestra da un total de 277, 74 redondeado a 278 personas.

Para poder tener una muestra que sea homogénea a la población la técnica de muestreo que se utiliza es la de por conglomerados por como su nombre indica esta técnica trata de escoger de forma aleatoria ciertos barrios o conglomerados dentro de una región, ciudad, comuna, para luego elegir unidades más pequeñas como escuelas, consultorios, hogares, etc. (Otzen & Monterola, 2017)

Al usar la técnica de muestro por conglomerados se tendrá que escoger conglomerados de grupos para obtener una muestra homogénea, por lo cual a cada parroquia perteneciente a las 53 de Guayaquil que muestra la información de la Ecu911, se encuestara a 5 personas así obteniendo una muestra que pueda representar de manera más exacta a la población de la ciudad de Guayaquil.

## **CAPÍTULO IV**

### **PROPUESTA TECNOLÓGICA**

Para el modelo predictivo que analice cuales son las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y a partir de esta información predecir qué sectores en desarrollo tendrán el mismo nivel de delincuencia, por lo cual en este capítulo se presentara los pasos que se siguieron para lograr obtener un modelo predictivo describiendo las herramientas utilizadas, la técnica de minería de datos escogida y la metodología, por lo cual se hizo diferentes comparaciones entre distintas herramientas a través de la técnica de benchamark.

#### **HERRAMIENTAS TECNOLÓGICAS UTILIZADAS**

##### **Herramienta de desarrollo**

La herramienta en la cual se va a implementar la metodología de KDD es Anaconda Jupyter, la cual es un entorno web de desarrollo interactivo, el cual permite a usuarios configurar y organizar flujos de trabajos, ciencia de datos, computación científica, periodismo computacional y hasta machine learning gracias a su interfaz flexible y extensiones para enriquecer su funcionalidad. (Jupyter, s.f)

La razón por la que se escogió usar Jupyter fue tras realizar la técnica de benchamark, la cual compara varias herramientas entre sí para determinar la más óptima para el proyecto de titulación.

**Tabla 2***Tabla benchmark de herramientas de desarrollo*

|                          | Jupyter   | Weka  | Knime   |
|--------------------------|---|---|---|
| Ventajas                 | Funciona en el navegador<br>Código en vivo (Live-Code)<br>Diferentes opciones a la hora de exportar y compartir los resultados<br>Control de versiones<br>Permite colaboración (JupyterHub)<br>Soporta más de 50 lenguajes de programación. (IONOS, ionos.es, 2019) | Contiene una gran gama de técnicas para modelado y procesamiento de datos<br>Weka puede funcionar en casi todas de las plataformas actuales al ser una implementación en java (Cordoba, 2011) | Comunidad proactiva.<br>Se integran nuevos desarrollos continuamente.<br>Flujos muy intuitivos con componentes que pueden ser reutilizados. (Andalucía, 2022) |
| Lenguaje de programación | <b>Lenguaje por defecto: Python</b> / C++, R, Julia, Ruby, JavaScript, CoffeeScript, PHP, Java (IONOS, ionos.es, 2019)  | Java  | Java  |
| Sistema operativo        | Windows, macOS, Linux   | Windows, macOS, Linux   | Windows, macOS, Linux   |
| Precio/Licencia          | Gratis  | Gratis  | Software libre a partir de la version 2.1 (IONOS, ionos.es, 2022)   |

*Nota: Tabla desarrollada por el autor*

Como se puede observar en la tabla comparativa benchmark tras comparar tres de las herramientas de minería de datos más utilizadas, se puede llegar a la conclusión que la herramienta que presenta más facilidades de uso es el entorno web JupyterLab al presentar ventajas bastantes llamativas, una amplia gama de lenguajes de programación, compatible con varios sistemas operativos y con una licencia gratis que brinda un fácil acceso a los usuarios.

## Lenguaje de desarrollo

Para el desarrollo de un modelo predictivo es necesario saber que lenguaje de desarrollo tiene las características que sean más adecuadas para la realización de esta tarea que se usará en la herramienta de desarrollo escogida anteriormente la cual va a ser seleccionada después de realizar una comparación benchmark entre distintos lenguajes que JupyterLab soporte.

**Tabla 3**

*Tabla benchmark de lenguajes de programación*

|                    | Python   | Java  | R   |
|--------------------|--|---|---|
| Ventajas           | <p>Es un lenguaje fácil para los principiantes</p> <p>Python es multipropósito al no limitarse en trabajar dentro de la comunidad de ciencias de datos.</p> <p>Permite crecer y escalar junto a los proyectos al ser más rápido que R. (Edx, 2021)</p> | <p>Java es muy útil al desarrollar códigos de ETL y algoritmos de machine learning. (Morales, 2019)</p> | <p>El lenguaje R es un lenguaje que se usa bastante en el análisis de datos y la estadística. (Edx, 2021)</p> |
| Paradigma          | Multiparadigma   | Orientado a objetos   | Funcional y orientado a objetos   |
| Licencias          | Gratis   | Gratis  | Gratis  |
| Sistema operativos | Windows, macOS, Linux  | Windows, macOS, Linux   | Windows, macOS, Linux   |

*Nota: Tabla desarrollada por autor*

Después de realizar la comparación benchmark de los lenguajes de programación se puede notar las ventajas que presenta Python ante Java y R en el campo de las ciencias de datos al ser de uso adecuado para usuarios principiantes, al no encasillarse en un solo ámbito de las ciencias de datos y en cómo puede crecer y escalar más rápido que el lenguaje R. Presenta un Multiparadigma junto con su licencia gratis y soporte en todos los sistemas

operativos, además de poseer una amplia librería de paquetes como pandas lo cual lo hace adecuado para aplicaciones de aprendizaje automático.

### Metodologías de minería de datos

La minería de datos tiene distintas metodologías las cuales se aplican en una herramienta para la predicción de situaciones a partir de un set de datos inicial, el cual tras aplicar una técnica de minería de datos puede arrojar un resultado. Para escoger la mejor metodología que se usara, las tres metodologías principales se compararan y se escogerá la metodología más adecuada para la tarea a través de un benchmark.

**Tabla 4**

*Comparación benchmark entre metodologías de minería de datos*

|          | KDD                         | Crisp-Dm                 | SEMMA                                |
|----------|-----------------------------|--------------------------|--------------------------------------|
| Técnicas | Asociación                  | Asociación               | métodos estadísticos de agrupamiento |
|          | Agrupamiento                | Agrupamiento             | redes neuronales                     |
|          | Clasificación y predicción  | Clasificación            | árboles de decisión                  |
|          | Aprendizaje autónomo        |                          | reglas de asociación                 |
|          | Clustering                  |                          |                                      |
|          | reglas de asociación        |                          |                                      |
|          | arboles de decisión         |                          |                                      |
|          | naives bayesianos           |                          |                                      |
|          | reglas de asociación        |                          |                                      |
| Procesos | Selección de datos          | Comprensión del negocio  | Datos                                |
|          | Preprocesamiento y limpieza | Comprensión de los datos | Exploración                          |
|          | Transformación              | Modelamiento de datos    | Transformación                       |
|          | Minería de datos            | Modelado                 | Modelo                               |
|          | Interpretación              | Evaluación               | Evaluación                           |
|          |                             | Despliegue               |                                      |

*Nota: Tabla desarrollada por autor*

Como se observa en la tabla de las metodologías aplicadas a la minería de datos, la metodología KDD y SEMMA tienen procesos muy similares entre ellos con la diferencia que SEMMA está enfocado más al uso ligado a productos SAS mientras KDD se enfoca más en patrones de datos por lo cual será la metodología que se empleará y Crisp-DM se enfoca más en objetivos empresariales al tener pasos que comprenden el negocio en donde se aplicara el método.

### Técnica de minería de datos

En el proceso de ciencia de datos existen muchas técnicas empleadas para la minería de datos, pero de todas estas técnicas existentes se escogerá tres del grupo de técnicas de predicción ya que este proyecto de titulación va a diseñar un modelo predictivo por lo cual no será necesario comparar las demás técnicas existentes.

**Tabla 5**

*Comparación benchmark entre algoritmos de machine learning*

|                 | Random Forest  | Soporte vectorial   | Regresión   |
|-----------------|--|---|---|
| Características | El algoritmo Random Forest es un algoritmo perteneciente al aprendizaje supervisado y se destaca por generar múltiples árboles de decisión en un solo conjunto de datos de entrenamiento. Los resultados arrojados se fusionan con fin de obtener un modelo único más óptimo en comparación con los resultados de cada árbol de decisión individualmente. (Espinosa, 2020) | También son conocidas con el acrónimo SVM por sus siglas en inglés (Support Vector Machines) y en español soporte de máquina de vectores. Es un algoritmo de regresión y clasificación . (Martinez, 2019) | La regresión lineal es una técnica que predice valores desconocidos a través del uso de otro valor de datos históricos, modelando matemáticamente la variable de entrada y salida como una ecuación lineal. (Ammar, 2022) |

*Nota: Tabla desarrollada por autor*

Como se puede observar, las tres técnicas de machine learning comparten un rasgo en común, este es el hecho de que todas estas técnicas se las puede usar en el aprendizaje supervisado es decir que estas técnicas

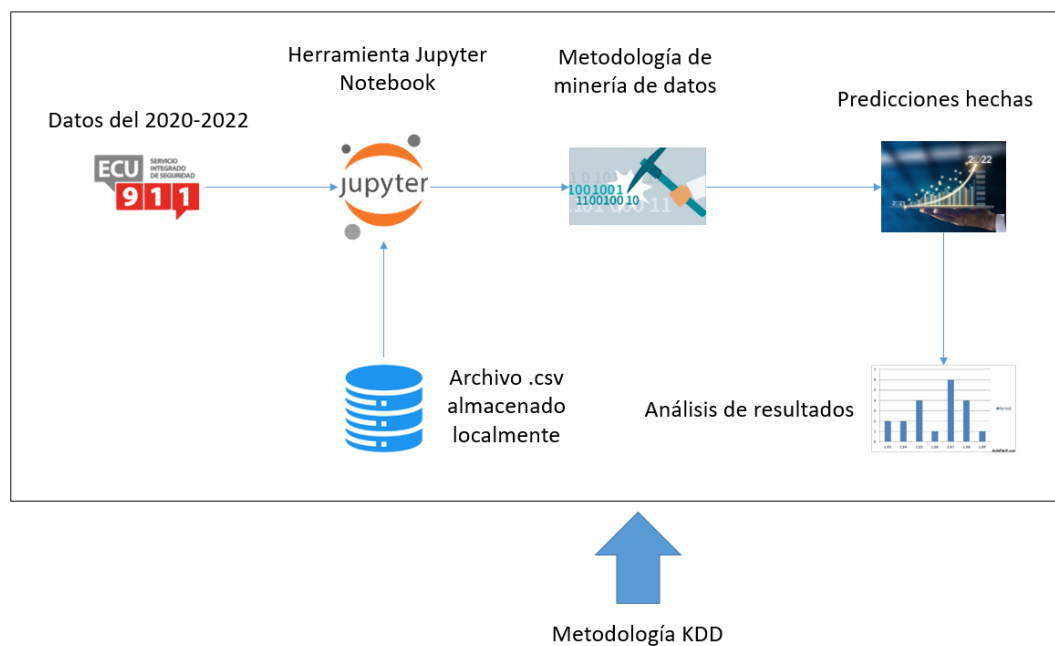


generan modelos que aprenden automáticamente dependiendo de las variables que se declaren como entrada y salida, por lo cual se decidió utilizar los tres algoritmos para ver como el modelo se comporta cuando se le aplica diferentes algoritmos de regresión.

## DESARROLLO DEL DISEÑO DE MODELO PREDICTIVO

### Diagrama del modelo predictivo usando KDD

**Figura 3**



*Nota: Figura hecha por autor donde se describe el diagrama de flujo*

El diseño del modelo predictivo usando la metodología Knowledge Discovery in Databases o también conocido como KDD esta simplificado de una manera abreviada para un entendimiento más óptimo.

Primero se obtienen los datos de las ocasiones en que personas de la ciudad de Guayaquil acudieron al Ecu911 por emergencias de seguridad ciudadana comprendido en los años 2020, 2021 y 2022. Luego se exporta

estos datos a la herramienta principal llamada Jupyter Notebook la cual es una herramienta online en donde los archivos creados se almacenan localmente en la computadora que se usa. Tras realizar la limpieza de datos y preprocesamiento se procede a ejecutar los algoritmos de machine learning escogidos para posteriormente obtener resultados de las predicciones que se desea llegar para finalmente usar todos los resultados obtenidos para compararlos y analizar su proximidad en la predicción.

Todo el proceso engloba a la metodología KDD por lo cual en el diagrama se muestra que, desde la obtención de datos, hasta el análisis de resultados, están rodeados por un rectángulo que resume el proceso de KDD.

### **Obtención de Dataframe**

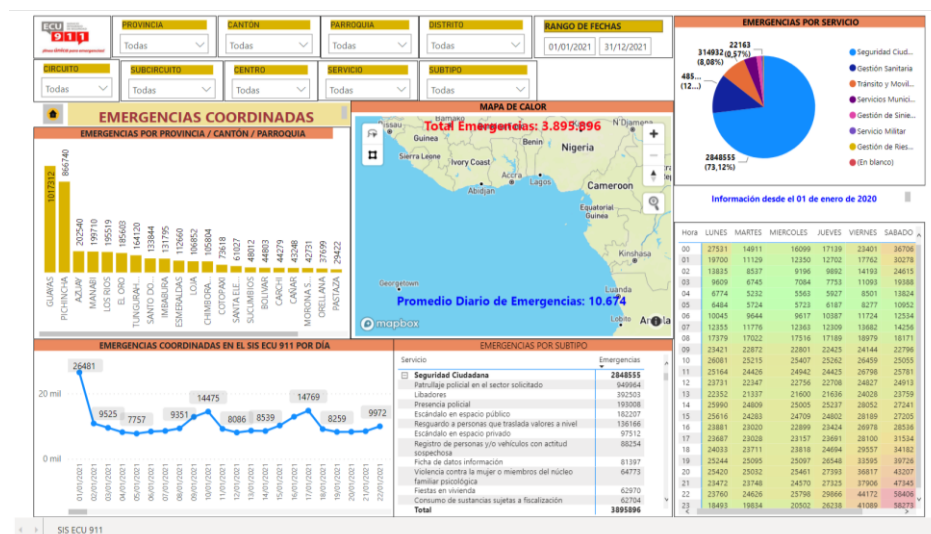
Para aplicar la metodología KDD es necesario extraer de alguna fuente un set de datos para posteriormente aplicar los pasos a seguir y finalmente conseguir diseñar el modelo predictivo. Para la extracción de datos confiables fue necesario la investigación de estos datos a instituciones del país que estén muchas veces en contacto con la delincuencia de la ciudad.

Las principales instituciones que se optó por buscar fue la fiscalía ya que en su página web muestra estadísticas de robos y femicidios con opción a pedir información más precisa al enviarles un correo electrónico con un plazo de espera de dos semanas. Al no recibir ninguna contestación de la fiscalía tras varios intentos se decidió investigar en otras instituciones como la policía para comprobar si brindaban información pública, pero no se encontró mayor información que noticias.

Finalmente se optó por buscar al Ecu911 los cuales si tienen un portal web el cual muestra públicamente todas las alertas que reciben por parte de los ciudadanos representado en gráficos de PoweBI donde se puede filtrar la información que uno desee en los diferentes tipos de emergencia que esta institución brinda. De todos los tipos de emergencias, la emergencia que sirve es la de Seguridad Ciudadana la cual contiene todos los servicios con la seguridad de los ciudadanos.

**Figura 4**

*Portal de la Ecu911*



*Nota: Figura obtenida del portal de la Ecu911*

Como se observa el usuario puede filtrar los datos por provincia, cantón, parroquia, distrito, rango de fechas, circuito, subcircuito, centro, servicio y subtipo de servicio. Con cada valor filtrado, las tablas y gráficos de PowerBI se actualizarán filtrando zonas de la ciudad que no son necesarios para el modelo predictivo como por ejemplo filtrar solo las parroquias urbanas de Guayaquil excluyendo las parroquias rurales que son Morro, Posorja, Puna y Tenguel.

## **Variables que determinan a un sector como peligroso**

Tras filtrar los datos de la Ecu911 para que solo muestre las parroquias urbanas del cantón de Guayaquil, se procede a realizar un análisis de las variables que constan dentro del servicio de seguridad ciudadana que puedan calificar como peligroso para los sectores de Guayaquil.

Para seleccionar que variable o situación consideraría a un sector como peligroso se hizo una limpieza manual de los subtipos del servicio de seguridad ciudadana ya que muchos de los subtipos de emergencias representan catástrofes o situaciones no relacionadas con la delincuencia y peligro de Guayaquil como es el caso de sismos, fuegos artificiales, accidentes, inundaciones, ruidos fuertes, entre muchas otras más que serán listadas a continuación.

Primero se mostrarán las variables que se clasificaron como no peligrosas al no representar ninguna relación con la delincuencia que atormenta a la ciudad de Guayaquil. Después se mostrarán las variables que serán consideradas para el set de datos al relacionarse con la delincuencia.

**Tabla 6***Variables categorizadas como no peligrosas*

| <b>Variables no Peligrosas</b>  |
|---|
| Seguridad en espectáculos públicos  |
| Encargos domiciliarios  |
| Vehículos con escapes modificados   |
| Vehículos de perifoneo  |
| Manejo de material pirotécnico  |
| Tráfico de fauna  |
| Comercialización de material pirotécnico  |
| Control campaña electoral   |
| Carreras de vehículos clandestinas  |
| Falso funcionario   |
| Desalojos   |
| Falsificación y uso de documento falso  |
| Maltrato a animales   |
| Aprovechamiento ilícito de servicios públicos                                   |
| Tráfico de flora  |
| Cierre de vías  |
| Moneda falsa  |
| Control de huelgas  |
| Pelear o combates entre perros  |
| Persona con dispositivo electrónico que abandona el rango territorial dispuesto |
| Riña interior de CRS  |
| Ruptura y/o abandono del dispositivo electrónico                                |
| Colaboración en desastres naturales   |
| Elaboración de material pirotécnico   |
| Falsificación de firmas   |

|   |
|---|
| Bullying a niños, niñas y adolescentes              |
| Causas naturales                                    |
| Gestión de Riesgos                                  |
| La adopción ilegal de niñas, niños y adolescentes.  |
| Trabajos forzados u otras formas de explotación     |
| Vehículos de gas                                    |
| Recuperación de animales silvestres                 |
| Organismos de tránsito                              |
| Manifestaciones                                     |
| Rescate y retención de especies silvestres          |
| Agentes municipales                                 |
| Descarga del dispositivo electrónico                |
| Control de manifestaciones                          |
| Organismos de salud                                 |
| Pirotécnia  |
| Boleta / orden de autoridad                         |
| Control y recuperación de flora y fauna             |
| Control de marchas                                  |
| Ocupación, uso ilegal de suelo o tráfico de tierras |
| Eventos clandestinos                                |
| Apoyo a instituciones articuladas                   |
| Eventos cívicos                                     |

*Nota: Tabla que contiene las variables categorizadas como no peligrosas que serán descartadas*

## Tabla 7

### Tabla de variables categorizadas como peligrosas

| Variables Peligrosas   |
|--|
| Abigeato   |
| Abuso de confianza   |
| Acercamiento agresor-víctima con dispositivo electrónico           |
| Acoso sexual   |
| Actos Inmorales  |
| Actos inmorales en la vía pública                                  |
| Agresión a la autoridad  |
| Agresión física  |
| Agresión verbal  |
| Agresiones a personas  |
| Alteración del dispositivo electrónico                             |
| Amenaza de bomba   |
| Apoyo al control de dispositivos electrónicos de carácter judicial |
| Asesinato  |
| Boleta de apremio  |
| Boleta de auxilio  |
| Boleta de captura  |
| Boleta de comparecencia inmediata                                  |
| Boleta de encarcelamiento  |
| Capturado por civiles  |
| Comercialización de sustancias sujetas a fiscalización             |
| Constatar persona sin vida   |
| Consumo de sustancias sujetas a fiscalización                      |
| Contrabando  |
| Custodia policial en arrestos domiciliarios                        |
| Custodia policial en casa de salud                                 |
| Custodia policial en lugares de riesgo temporal                    |
| Daño a los bienes policiales                                       |
| Daño a propiedad privada   |
| Daño a propiedad pública   |
| Daño a propiedad pública o privada                                 |
| Daño a señaléticas   |
| Delito hidrocarburos   |
| Delitos sexuales   |
| Desaparecido   |
| Desaparición de persona  |
| Desaparición forzada   |
| Disparos   |
| Diversas formas de explotación                                     |
| Estafa   |
| Estupro  |
| Evasión de arresto domiciliario                                    |
| Evasión de centro de detención                                     |
| Eventos ilegales   |
| Extorsión  |
| Extraviado   |
| Falsificación de moneda o documentos                               |
| Falta contra la integridad a servidores policiales                 |
| Falta contra la integridad a servidores públicos                   |
| Femicidio  |
| Ficha de datos   |
| Ficha de datos información   |

|  |
|--|
| Ficha de datos operativo   |
| Fraude a persona   |
| Homicidio  |
| Hurto  |
| Ingreso a escenarios de concurrencia masiva armas blancas, petardos, bengalas o material pirotécnico prohibido |
| La explotación sexual de niñas, niños y adolescentes   |
| Libadores  |
| Maltrato o muerte de mascotas o animales de compañía   |
| Muerte indeterminada   |
| Orden de incautación/embargo   |
| Osamentas  |
| Patrullaje policial en el sector solicitado  |
| Persona armada   |
| Persona con dispositivo electrónico no contactada  |
| Persona herida   |
| Persona herida con arma blanca   |
| Persona herida con arma de fuego   |
| Persona herida con objeto contundente  |
| Plantones  |
| Presencia policial   |
| Privación arbitraria de la libertad por civiles (Secuestro)  |
| Prostitución   |
| Receptación / cachinería   |
| Registro de personas y/o vehículos con actitud sospechosa  |
| Resguardo a personas que traslada valores a nivel local  |
| Resguardo a personas que traslada valores a nivel provincial   |
| Resguardo de valores de blindado   |

|  |
|--|
| Resguardo de víctimas y testigos                       |
| Resguardo policial                                     |
| Robo   |
| Robo a carros  |
| Robo a domicilio                                       |
| Robo a embarcaciones en espacios acuáticos             |
| Robo a entidades financieras                           |
| Robo a instituciones de salud                          |
| Robo a instituciones públicas                          |
| Robo a unidades económicas                             |
| Robo a unidades educativas                             |
| Robo a vehículos de transporte de valores              |
| Robo accesorios de vehículos o autopartes de vehículo  |
| Robo de bienes patrimoniales                           |
| Robo de motos  |
| Robo en ejes viales o carreteras                       |
| Robo personas  |
| Secuestro  |
| Secuestro extorsivo                                    |
| Seguridad de autoridades                               |
| Sicariato  |
| Sonidos de alarma                                      |
| Sustracción de bienes del patrimonio cultural          |
| Tenencia ilícita de sustancias sujetas a fiscalización |
| Tenencia y porte de arma blanca o cortopunzante        |
| Tenencia y porte de armas de fuego                     |
| Tenencia y porte de explosivos                         |
| Tentativa de robo                                      |
| Tentativa de secuestro                                 |
| Tráfico de moneda                                      |

|   |
|---|
| Traslado de detenidos   |
| Traslado de valores   |
| Trata de personas   |
| Usurpación de uniformes e insignias                                       |
| Venta de bebidas alcohólicas a menores de edad                            |
| Venta u ofrecimiento de bebidas alcohólicas o cigarrillos menores de edad |
| Violación   |
| Violencia a niños, niñas y adolescentes                                   |
| Violencia contra la mujer o miembros del núcleo familiar física           |
| Violencia contra la mujer o miembros del núcleo familiar psicológica      |
| Violencia contra la mujer o miembros del núcleo familiar sexual           |
| Violencia intrafamiliar   |

*Nota: Variables categorizadas como peligrosas que serán consideradas para el set de datos obtenidas del portal del Ecu911.*

Como se puede observar de todas las variables pertenecientes al servicio de seguridad ciudadana, la mayor parte pertenece a la categoría de peligroso por eso esas variables se las considerara como parte de las emergencias que definirían a un sector como peligros las cuales han sido reportadas por las personas de Guayaquil.

### **Preprocesamiento y limpieza de datos**

Después de realizar la limpieza y filtrado manual de las variables o emergencias pertenecientes al servicio de seguridad ciudadana que determinan a un sector como peligroso, se procede a descargar todas las tablas y gráficos en formato .csv para posteriormente importarlos a la herramienta de Jupyter Notebook y poder realizar la limpieza de datos.

Al obtener los datos principalmente de una institución que comparte públicamente sus datos, es de esperar que sus datos ya hayan pasado por un proceso interno de limpieza y transformación, por lo que no tendría que ser necesario la limpieza y preprocesamiento de esos datos al ya estar limpios, pero de igual manera se procederá a realizar un proceso de limpieza para demostrar que no hay datos nulos, corruptos, repetidos o erróneos.



En la herramienta de Jupyter Notebook se importa la librería pandas la cual es una librería de Python que permite leer y escribir ficheros en formato csv, Excel y base de datos SQL, ofrece métodos para reordenar, dividir y combinar conjuntos de datos realizando todas estas operaciones de manera muy eficiente. Pandas también proporciona tres estructuras de datos diferentes los cuales son series, DataFrames (tablas) y Panels (cubos). (Sanchez, 2022)

### **Figura 5**

*Paquete de pandas*

```
import pandas as pd
```

*Nota: figura hecha por autor*

Tras importar pandas se procede a mostrar la tabla de contenido con el método head(53) el cual muestra los 53 registros de la tabla de datos del número de emergencias que han sido reportadas desde el 01/01/2020 hasta el 31/12/2020 con sus respectivas zonas mostrando en total cuatro columnas con 52 registros cada una. La columna DPA\_DESPRO refiere a la provincia por la cual se filtró la información en el portal web de la Ecu911, DPA\_DESCAN refieren al cantón Guayaquil, DPA\_DESP contiene a la información de parroquias y sectores de Guayaquil y finalmente StatusIncident2020, StatusIncident2021 y StatusIncident2022 que son el número de emergencias registradas en los años 2020, 2021 y 2022 de los sectores y parroquias.

## Figura 6

Set de datos obtenidos del Ecu911

```
df.head(53)
```

|   | DPA_DESPRO | DPA_DESCAN | DPA_DESP                   | StatusIncident2020 | StatusIncident2021 | StatusIncident2022 |
|---|------------|------------|----------------------------|--------------------|--------------------|--------------------|
| 0 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-PORTETE          | 12468              | 16312              | 15711              |
| 1 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL- EL CISNE        | 12086              | 15839              | 12877              |
| 2 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-GARAY            | 11932              | 16989              | 15298              |
| 3 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-GUASMO           | 10897              | 13840              | 9999               |
| 4 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-SALINAS          | 9979               | 12567              | 11794              |
| 5 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-NUEVA PROSPERINA | 9366               | 11495              | 10107              |

Nota: Figura hecha por autor

Se realiza el método de info() el cual mostrara la informacion del set de datos como sus variables numéricas y variables categóricas, igualmente como su número de registros y la cantidad de valores no nulos.

## Figura 7

Datos que muestra nos tipos de datos que contiene el dataframe

```
print(df.info())  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 53 entries, 0 to 52  
Data columns (total 6 columns):  
#   Column                Non-Null Count  Dtype  
---  ---                -  
0   DPA_DESPRO            53 non-null     object  
1   DPA_DESCAN            53 non-null     object  
2   DPA_DESP              53 non-null     object  
3   StatusIncident2020    53 non-null     int64  
4   StatusIncident2021    53 non-null     int64  
5   StatusIncident2022    53 non-null     int64  
dtypes: int64(3), object(3)  
memory usage: 2.6+ KB  
None
```

Nota: Figura desarrollada por autor

Ya desde este punto se puede comenzar a verificar que no se encuentra ningún registro que contenga valores nulos, aun así, se procederá a realizar sus respectivas funciones para confirmar lo mencionado recientemente.

La función perteneciente a pandas de `isna()` es un método que permite identificar valores faltantes o valores Nan, por lo que se puede observar que se realizó un conteo para contar el número de variables vacías que existen en el dataframe y al ver lo que bota se puede observar que no existen valores vacíos.

### Figura 8

Figura que muestra que no hay valores nan en la tabla datos

```
counts = df.isna().sum()
print(counts.sort_values())
```

```
DPA_DESPRO          0
DPA_DESCAN          0
DPA_DESP            0
StatusIncident2020  0
StatusIncident2021  0
StatusIncident2022  0
dtype: int64
```

Nota: Figura diseñada por autor

Tras verificar que no existen valores nulos en el set de datos, lo siguiente que se realizó fue la eliminación de registros que contenían valores nulos. El set de datos al no tener datos nulos, no debería de eliminarse ningún registro al ser un dataframe ya limpio, de igual manera se implementó el método de `dropna()` para eliminar valores nulos o nan.

### Figura 9

Dataframe sin valores nulos o vacíos

|    |        |           |                           |      |      |      |
|----|--------|-----------|---------------------------|------|------|------|
| 48 | GUAYAS | GUAYAQUIL | GUAYAQUIL-LOS CEIBOS      | 2221 | 3615 | 3721 |
| 49 | GUAYAS | GUAYAQUIL | GUAYAQUIL-NUEVO GUAYAQUIL | 2194 | 2620 | 2615 |
| 50 | GUAYAS | GUAYAQUIL | GUAYAQUIL-CHONGON         | 1167 | 1470 | 1570 |
| 51 | GUAYAS | GUAYAQUIL | GUAYAQUIL-PUERTO HONDO    | 932  | 1227 | 1631 |
| 52 | GUAYAS | GUAYAQUIL | GUAYAQUIL-PUENTE LUCIA    | 862  | 754  | 703  |

Nota: Figura hecha por autor

Como se puede observar en la imagen anterior en donde se muestran los últimos cinco registros, no se ve que se ha eliminado ningún registro dando

a entender que no existían datos vacíos o nulos que tengan que ser eliminados.

Tras el proceso de eliminación de valores nulos, sigue el proceso de eliminación de datos repetidos, para esto se usa la función `duplicated()` que devuelve en valor booleano, es decir de verdadero o falso si existe algún valor duplicado o repetido en el set de datos.

En los 52 registros se mostró un valor de false dando a entender que no existen valores en ninguno de los 52 registros algún valor repetido por lo cual se procede a ejecutar la función de `drop_duplicates()` que elimina registros con valores duplicados y al no tener ningún valor duplicado devolverá los 52 registros.

Con estas pruebas realizadas se puede concluir que en verdad los datos obtenidos del portal web de la Ecu911 ya eran registros que habían sido pasados por un proceso previo de limpieza antes de liberarlos al público para que puedan ser visualizados por cualquier persona.

### **Técnica de minería de datos**

Para esta etapa de KDD se realizarán tres corridas en total, cada corrida con una técnica de machine learning diferente junto a su probabilidad de exactitud, su error medio cuadrático y su error medio absoluto. Para cada técnica se aplicará una variable  $X$  para el entrenamiento del modelo y para su prueba se le asignará la variable  $y$ . Con estas variables se entrenará el modelo dependiendo del algoritmo que se esté usando para finalmente predecir los valores futuros.

Tras la obtención de los valores futuros de la predicción, se procederá a usar estos nuevos valores junto a los datos previamente usados para así predecir más valores futuros, este proceso se lo realizará unas 2 veces y tras hacerlo se usará los datos obtenidos para comparación y análisis de resultados.

## **Regresión lineal**

Una de las técnicas de minería de datos que se aplicó en el set de datos es la regresión lineal la cual pertenece al paquete de scikit-learn llamado `LinearRegression()` que es una técnica de machine learning que corresponde al aprendizaje supervisado al tener conocimiento de los datos de entrada como los de salida.

Se importa las librerías de `numpy`, `pandas` y `train_test_split` que permite el entrenamiento del modelo. Tras esto se importa el archivo del set de datos `csv` y se comienza a preparar los datos de entrada y salida denominados `X` y `Y` para siguiente paso es entrenar el modelo con la regresión lineal y ejecutar el método `predict()` para obtener los valores de la predicción.

## Figura 10

### Código de la regresión lineal

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
df = pd.read_csv("Datos2020-22.csv")

# Preparar Los datos para el modelo
X = df[['StatusIncident2020', 'StatusIncident2021']]
y = df['StatusIncident2022']

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

LinearRegression()

# Hacer la predicción para 2023
poblacion_2023 = model.predict(X)
print("La población prevista para 2023 es:", poblacion_2023)

La población prevista para 2023 es: [13977.08768572 13587.16430467 14578.72567522 11922.15526659
10868.31461981 9974.43308254 11035.89686706 13706.01455416
11297.48265248 8512.96370205 7993.14896219 11595.71381869
8605.44436626 10664.05856507 11715.54300816 10550.37786067
8242.8382448 9128.09204871 6341.86155843 7644.92010801
7748.83030796 9486.28303771 7225.65968346 6815.95638815
7152.2159344 6678.82483346 7192.9086367 6470.55417116
8302.04277657 5594.66416688 5207.3592814 5910.21081107
7459.58176671 6373.69739386 6143.70519759 6402.19656939
6229.23064894 5068.32178542 5055.27475783 5737.41792368
5401.40301832 5519.46172847 3980.64317027 3702.99291056
4132.72303012 4041.90114624 6029.51144845 3612.92713855
3509.8681703 2658.39648525 1714.33098872 1515.56935985
1113.09838126]
```

*Nota: Código hecho por autor*

Tras obtener los resultados de la predicción del 2023 de las emergencias pertenecientes a las emergencias realizadas por las personas, se procede a verificar que tan confiable y preciso es el modelo usando regresión lineal. Para esto se importa los paquetes de `mean_squared_error`, `r2_score` y `mean_absolute_error` que corresponden a error cuadrático medio (MSE) que calcula como la media de los errores al cuadrado entre los valores reales y los valores predichos, `r2` se define como la proporción de la varianza en los valores reales explicada por la varianza en los valores predichos. Un  $R^2$  cercano a 1 indica una buena precisión de la predicción, mientras que un  $R^2$  cercano a 0 indica una precisión baja. y error absoluto medio (MAE) el cual calcula como la media de los errores absolutos entre los valores reales y los valores predichos.

## Figura 11

### Error y precisión de la predicción

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)
```

```
MSE: 1281428.7283904164
MAE: 752.0649487929595
r2: 0.891207458752686
```

*Nota: Código hecho por autor*

Se puede observar que el r2 se obtiene un valor de 0.8912, un valor muy acercado al 1 que como se mencionó anteriormente, mientras más cercano se encuentre al 1 más preciso será la predicción realizada. También presenta un error medio absoluto 752.06 y un error cuadrático medio de 1281428.72

Tras haber obtenido respuestas de predicciones asignando valores de entrenamiento los años 2020 y 2021 se decidió empezar a realizar pruebas para observar que pasaría si solo se usa los valores del 2020 como variables de entrada y los valores del 2022 como variables de salida y el resultado obtenido fue el siguiente.

## Figura 12

### Variables de entrada y salida

```
# Preparar Los datos para el modelo
X = df[['StatusIncident2020']]
y = df[['StatusIncident2022']]
```

*Nota: Código hecho por autor*

Se declaró X como la variable de entrada que contiene las emergencias solamente del año 2020 y la variable Y contiene las emergencias del 2022.

### Figura 13

*Variables de entrada y salida*

```
La población prevista para 2023 es: [14482.76080858 14059.7438259 13889.20818367 12743.07578295
11726.50617535 11047.68572932 10816.24450057 10574.83690313
10361.11366319 10307.95969678 10158.46416625 10147.39042325
9271.45735178 9022.29813423 8900.4869612 8778.67578818
8733.27344187 8682.33422406 8015.69489533 7950.35981162
7849.5887503 7737.74394598 7603.75165565 7583.81891825
7526.23545464 7514.05433734 7396.67266151 7326.9080806
7268.21724269 7261.57299689 6388.96204832 6382.31780252
6376.78093102 6336.91545621 6276.0098697 6142.01757937
5666.95400458 5569.50506616 5433.29802724 5076.72350257
5056.79076516 5014.71054175 4682.49825169 4541.86171556
4317.06473262 4246.19277741 4135.45534738 3199.7240637
3135.49635429 3105.59724819 1968.32384187 1708.09088132
1630.5746803 ]
MSE: 2435935.7782707713
MAE: 1241.802881297607
r2: 0.7931904929537456
```

*Nota: Código hecho del autor*

La respuesta obtenida de la predicción dio como resultado una predicción menos precisa que la anterior ya que se está usando solamente un a variable de entrada con un r2\_score de 0.7931, valor menor que cuando se usó de entrada los años 2020 y 2021.

Siguiente se usó la variable de entrada a los valores del año 2021 para observar que cambios hay en las predicciones.

### Figura 14

*Variables de entrada y salida de los años 2021 y 2022*

```
X = df[['StatusIncident2021']]
y = df[['StatusIncident2022']]
```

*Nota: Código hecho por autor*

Los resultados obtenidos del modelo al usar los datos del año 2021 son los siguientes.



## Figura 15

### Variables de entrada y salida

```
La población prevista para 2023 es: [14019.59890134 13627.05016114 14581.44984663 11968.05453393
10911.57557773 10021.90908831 11041.87188262 13619.58094621
11279.22693484 8580.35060684 8071.61407883 11560.56736387
8633.46502411 10618.61636993 11632.76977486 10499.93884382
8263.3239287 9118.96399456 6396.85010899 7656.65769384
7753.75748793 9432.67102161 7238.38165776 6840.85344094
7164.51942123 6705.57765943 7199.37575757 6497.26955416
8269.13331809 5646.60896492 5240.7816204 5921.31009179
7421.79237993 6368.63307481 6143.72671415 6389.38089406
6205.14025913 5077.28880471 5059.86063654 5708.0225099
5381.8667913 5494.73492802 3992.59281434 3718.72160024
4127.03868307 4036.57819115 5957.82625367 3584.2757315
3482.19646079 2656.43325465 1702.03356917 1500.36476606
1107.81602585]
MSE: 1282700.2819010115
MAE: 760.4109278196763
r2: 0.8910995046115897
```

*Nota: Código hecho por autor*

Se puede ver que en esta ocasión la predicción obtenida vuelve a tener una aproximación más estimada de 0.8910, un valor muy parecido que cuando se usó las variables de entrada de los años 2020 y 2021. Con los resultados de estas diferentes corridas hechas de los modelos predictivos con el algoritmo de regresión lineal, se decidió usar solamente una variable de entrada y una de salida ya que no es necesario cargar tantas variables de entrada para obtener una predicción óptima.

A continuación, se insertó los datos del 2023 en el dataset y a partir de este punto se volvió a aplicar el algoritmo de regresión lineal, pero en este caso solamente usando la variable de entrada como el año 2022 y la variable de salida el año 2023 que corresponde a la predicción anteriormente hecha.

## **Figura 16**

*Coeficiente de precisión y errores de la predicción*

```
MSE: 1142230.0531302511  
MAE: 724.3702670095591  
r2: 0.89119191391409
```

*Nota: Código hecho por autor*

Con los nuevos datos resultante de la predicción, se puede notar que el valor de `r2_score` sigue siendo de un 0.89 tal como fue en el ejemplo anterior cuando el dato de entrada era solo el año 2021.

Tras obtener predicciones del año 2024, se vuelve a hacer el mismo proceso de insertar en el dataframe las emergencias del 2024 para próximamente usar este set de datos para la predicción del año 2025.

## **Figura 17**

*Variables de entrada y salida*

```
MSE: 1273.9257966577877  
MAE: 28.151173072677246  
r2: 0.9998786342611244
```

*Nota: Código hecho por autor*

Como se ve en los resultados de la predicción obtenida, las emergencias para el 2025 tienen una precisión de casi 1 lo cual es algo optimo con un valor de 0.99987

## **Random Forest**

El algoritmo de aprendizaje supervisado random forest establece puntos de salida desde los resultados de las decisiones generada por cada uno de los árboles individualmente. Para poder predecir con este algoritmo es necesario hacer un promedio de todos los elementos de salida de los árboles.

Así, se obtiene un promedio con mayor cantidad de árboles, mejor serán los resultados. (KeepCoding, 2022)

Con la técnica de minería de datos de random forests, el proceso para la predicción de los datos es muy parecido a la de regresión lineal, ya que de igual manera se le asigna una variable X y Y para entrenamiento y pruebas para posteriormente entrenarlo y predecirlo.

### **Figura 18**

*Código de predicción con random forest*

```
# Hacer la predicción para 2023
poblacion_2023 = model.predict(X)
print("La población prevista para 2023 es:", poblacion_2023)
```

La población prevista para 2023 es: [14729.75 13062.02 14874.02 11349.82 12097.49 10592.33 13570.35 10509.07 11683.23 8052.97 7956.16 11322.86 8268.25 11631.59 12237.68 11408.83 8153.78 8254.48 6025. 7486.87 8183.06 8923.86 6497.08 6629.39 7485.03 6105.03 7253.46 5805.64 8461.53 5517.88 5016.29 4460.89 7543.88 7266.62 6806.6 6446.98 6322.42 4723.23 5031.07 5663.21 5944.33 6779.83 3994.23 3651.3 4557.44 5120.82 6151.67 3482.68 3461.36 2896.73 1537.35 1476.27 1015.32]

*Nota: Código hecho por autor*

Tras obtener las predicciones con el algoritmo de random forest se procede a verificar sus medias de errores y aproximación  $r2\_score$ .

### **Figura 19**

*Índice de error y precisión de random forest para el año 2023*

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)
```

MSE: 237298.26982452828  
MAE: 360.41075471698105  
r2: 0.9798535172219616

*Nota: Código hecho por autor*

En este caso con el algoritmo de random forest se obtiene un error medio absoluto más reducido de 360.41 y una aproximación estimada  $r2$  de

0.9798, muy cercano al valor 1 lo cual quiere decir que esta predicción tiene un índice de aproximación muy alto y finalmente un error cuadrático medio de 237298.267.

En el caso de random forest, también se probó el modelo con solamente una variable de entrada siendo esta 2021 y la variable de salida 2022, lo cual dio como resultado casi exactamente los mismos valores tanto en la predicción como en la eficiencia del modelo.

### **Figura 20**

*Índice de error y precisión de random forest para el año 2023 con una variable de entrada*

```
La población prevista para 2023 es: [14596.02 12407.22 15069.6 10938.52 12405.72 10162.46 13829.73 11018.94
12484.9 8029.44 7950.68 11434.63 8182.64 11517.51 12167.44 11180.83
8188.39 8310.93 6049.89 7459.92 8161.67 8754.99 6511.16 6704.1
7575.41 5987.09 7301.65 5706.28 8335.27 5766.31 5109.29 4360.79
7345.04 7301.29 6705.56 6145.09 6388.45 4689.4 4997.24 5505.33
6104.41 6810.55 4081.34 3577.03 4556.6 5118.93 5638.03 3567.67
3661.27 2896.73 1537.35 1562.97 1015.32]
MSE: 245410.9502830189
MAE: 370.05132075471687
r2: 0.9791647554485969
```

*Código hecho por el autor*

Se procederá a hacer lo mismo que con el algoritmo de la regresión lineal, se insertará estos valores en el set de datos para posteriormente usar estos nuevos valores para predecir las emergencias del año 2024 y realizar su predicción respectivamente.

### **Figura 21**

*Valores de la predicción junto con su índice de precisión*

```
La población prevista para 2024 es: [14538.97 12479.91 14610.6 10972.03 12173.16 10358.92 13753.3 11167.9
12178.97 8026.56 7963.05 11281.28 8230.19 11731.31 12128.46 11117.19
8173.89 8322.3 6037. 7322.93 8077.9 8687.04 6499.34 6822.81
7517.42 5959.9 7260.13 5769.78 8293.71 5799.59 5158.4 4556.51
7365.04 7416.14 6855.74 6149.27 6318.4 4778.33 5066.27 5461.57
5880.87 6844.16 4119.83 3702.59 4456.77 5270.84 5831.65 3584.09
3586.79 3084.88 1536.54 1515.72 1221.09]
MSE: 16793.9046245283
MAE: 95.11716981132079
r2: 0.9984775307239794
```

*Nota: Código hecho por autor*

Como se observa la predicción con los valores de entrada del año 2020 hasta el 2022 con una variable de salida 2023 para predecir las emergencias del año 2024 se obtuvo unos valores que según su precisión se acercan mucho al 1 dando a entender que la predicción tiene un gran índice de precisión.

De igual manera con estos datos obtenidos se vuelve a llenar el dataframe y se lo volverá a usar para conseguir datos del 2025 usando las variables de entrada de los años pasados.

## **Figura 22**

*Valores de la predicción junto con su índice de precisión*

```
La población prevista para 2025 es: [14338.27 12311.71 14402.15 10800.05 12105.47 10564.48 13809.22 11022.67
12157.66 7928.92 7860.2 11194.08 8113.27 11742.01 12181.56 11006.89
7988.67 8107.5 5941.49 7236.68 8209.97 8594.27 6286.72 6892.92
7525.37 5923.4 7218.89 5796.89 8335.8 5818.93 5152.89 4188.28
7477.93 7719.66 6932.1 6063.87 6268.43 4768.92 5115.87 5477.79
6008.96 6962.37 4040.89 3937.67 4564.59 5395.78 5854.03 3767.06
3767.4 3298.25 1558.21 1553.32 1367.13]
MSE: 6413.088307547182
MAE: 56.03566037735853
r2: 0.9993969715823626
```

*Nota: Código hecho por el autor*

Con la predicción obtenida mostrada anteriormente, se observa que el valor de `r2_score` sigue en el rango del 0.999 lo cual indica que para años futuros su predicción tendrá una proyección que tiene de igual manera una precisión que se acerca al 1.

## **Soporte de vectores**

SVM (máquina de vectores de soporte) es una técnica de clasificación y regresión que aprovecha al máximo la precisión de las predicciones de un modelo sin ajustar excesivamente los datos de entrenamiento. SVM es ideal para analizar datos con un gran número de campos de predictores (por ejemplo, miles). SVM tiene aplicaciones en multitud de disciplinas, incluyendo la gestión de relaciones con los clientes (CRM), el reconocimiento facial y de otras imágenes, bioinformática, extracción de conceptos de minería de texto,

detección de intrusiones, predicción de estructura de proteínas y reconocimiento de la voz. (IBM, ibm.com, 2021)

Para aplicar este algoritmo se vuelve de la misma manera a importar sus librerías y paquetes, se asigna variables de entrenamiento y prueba, se entrena el modelo y finalmente se aplica la predicción para como último paso observar que tan acertadas han sido las predicciones obtenidas.

### Figura 23

*Predicción hecha con SVR*

```
# Hacer la predicción para 2023
poblacion_2023 = model.predict(X)
print("La población prevista para 2023 es:", poblacion_2023)
```

La población prevista para 2023 es: [7034.81020184 7035.82541943 7034.54161413 7039.31157271 7040.68695842 7040.29829987 7041.22051197 7037.74413585 7041.21711772 7036.48173239 7034.45271154 7041.03786951 7035.6593723 7040.17147608 7040.03093095 7039.7546032 7033.31533462 7036.5448493 7023.34323602 7028.99999998 7029.2470127 7035.82567544 7026.23851485 7024.31984493 7025.71376455 7023.54650837 7025.59482437 7022.20773337 7030.39364253 7018.83591704 7015.95461734 7017.98936161 7024.51748888 7019.63756271 7018.62708692 7019.38010619 7017.91158555 7014.40382113 7014.22337523 7015.55072053 7014.62610006 7014.89077689 7012.68825355 7012.71458897 7012.65366724 7012.64180047 7015.93803614 7013.27560627 7013.39291017 7014.37483162 7017.2490489 7017.94078132 7018.8229855 ]

*Nota: Código hecho por autor*

Se puede observar que tras realizar la predicción, sus valores que corresponderían a las emergencias del 2023, todos sus valores de predicción rondan en los valores de 7000, por lo cual en obvias razones se puede observar que este algoritmo no aplica para este escenario en específico. Esto se puede confirmar con los valores de los errores y aproximación r2.

### Figura 24

*Valores de la predicción junto con su índice de precisión de SVR*

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)
```

MSE: 11847208.960115278  
MAE: 2666.3193505549093  
r2: -0.005821034680439485

*Nota: Código hecho por autor*

Se puede observar que el valor que más destaca entre los tres es el  $r^2$  score que como se recordara este valor tiene que estar entre 0 y 1 pero en este caso se encuentra en los valores de los negativos representando que el modelo no es útil para predecir valores futuros.

La razón por la que soporte de máquina de vectores no aplica para este caso en específico a pesar que este algoritmo se lo puede usar en regresiones es porque el uso primario de las Máquinas de Soporte de Vectores es en escenarios de clasificación binaria al separar los datos en dos categorías de valores binarios, es decir en dos grupos diferentes. (Sotaquirá, 2021)

### Zonas con mayor índice de peligrosidad

Después de obtener las predicciones de las emergencias desde el 2023 hasta el 2025 se procede a identificar qué zonas de la ciudad de Guayaquil son las que han tenido un mayor crecimiento de emergencias reportadas a lo largo de la predicción realizada, para obtener las zonas que muestren el mayor índice de peligrosidad se realizó una suma de todas las emergencias de todos los años en cada zona y se mostró las 10 primeras zonas que presentaron un mayor número de emergencias.

### Figura 25

Tabla con las zonas que presentaron mayor crecimiento en la delincuencia en los datos regresión lineal

|        |                                 |       |       |       |       |       |       |       |
|--------|---------------------------------|-------|-------|-------|-------|-------|-------|-------|
| GUAYAS | GUAYAQU GUAYAQUIL-GARAY         | 11932 | 16989 | 15298 | 14578 | 14581 | 14577 | 87955 |
| GUAYAS | GUAYAQU GUAYAQUIL-PORTETE       | 12468 | 16312 | 15711 | 13977 | 14019 | 13976 | 86463 |
| GUAYAS | GUAYAQU GUAYAQUIL- EL CISNE     | 12086 | 15839 | 12877 | 13587 | 13626 | 13586 | 81601 |
| GUAYAS | GUAYAQU GUAYAQUIL- Esteros      | 8939  | 15830 | 9859  | 13706 | 13619 | 13705 | 75658 |
| GUAYAS | GUAYAQU GUAYAQUIL-GUASMO        | 10897 | 13840 | 9999  | 11922 | 11967 | 11921 | 70546 |
| GUAYAS | GUAYAQU GUAYAQUIL-MARTHA ROLDOS | 9157  | 12724 | 14975 | 11035 | 11041 | 11034 | 69966 |
| GUAYAS | GUAYAQU GUAYAQUIL-ATARAZANA     | 7427  | 13436 | 13205 | 11715 | 11632 | 11714 | 69129 |
| GUAYAS | GUAYAQU GUAYAQUIL-SUBURBIO      | 8553  | 13349 | 10762 | 11595 | 11560 | 11594 | 67413 |
| GUAYAS | GUAYAQU GUAYAQUIL-SAUCES        | 8746  | 13010 | 11527 | 11297 | 11278 | 11296 | 67154 |
| GUAYAS | GUAYAQU GUAYAQUIL-SALINAS       | 9979  | 12567 | 11794 | 10868 | 10911 | 10867 | 66986 |

Nota: En esta tabla se muestra la suma de todas las emergencias realizadas y predichas para obtener un crecimiento aproximado.

## Figura 26

Tabla con las zonas que presentaron mayor crecimiento en la delincuencia en los datos de random forest

|                         |       |       |       |       |       |          |          |
|-------------------------|-------|-------|-------|-------|-------|----------|----------|
| GUAYAQUIL-GARAY         | 11932 | 16989 | 15298 | 15069 | 14602 | 14402,15 | 88292,15 |
| GUAYAQUIL-PORTETE       | 12468 | 16312 | 15711 | 14596 | 14526 | 14338,27 | 87951,27 |
| GUAYAQUIL-MARTHA ROLDOS | 9157  | 12724 | 14975 | 13829 | 14056 | 13809,22 | 78550,22 |
| GUAYAQUIL- EL CISNE     | 12086 | 15839 | 12877 | 12407 | 12299 | 12311,71 | 77819,71 |
| GUAYAQUIL-SALINAS       | 9979  | 12567 | 11794 | 12405 | 12172 | 12105,47 | 71022,47 |
| GUAYAQUIL-ATARAZANA     | 7427  | 13436 | 13205 | 12167 | 12224 | 12181,56 | 70640,56 |
| GUAYAQUIL-SAUCES        | 8746  | 13010 | 11527 | 12484 | 12197 | 12157,66 | 70121,66 |
| GUAYAQUIL- Esteros      | 8939  | 15830 | 9859  | 11018 | 10914 | 11022,67 | 67582,67 |
| GUAYAQUIL-GUASMO        | 10897 | 13840 | 9999  | 10938 | 10742 | 10800,05 | 67216,05 |
| GUAYAQUIL-SUBURBIO      | 8553  | 13349 | 10762 | 11434 | 11173 | 11194,08 | 66465,08 |

Nota: En esta tabla se muestra la suma de todos las emergencias realizadas y predichas para obtener un crecimiento aproximado del random forest.

Se puede observar en ambas tablas como a pesar que los órdenes de las 10 primeras zonas siguen siendo las mismas, sus posiciones no lo son indicando que en conclusión que estos 10 sectores corresponden a las zonas más peligrosas de la ciudad al presentar más llamados de emergencia.

### Determinar qué zonas se volverán más peligrosas

Ahora toca determinar qué zonas serán las que se vuelvan más peligrosas en el tiempo, para esto se decidió sacar el incremento porcentual de cada año con la fórmula de incremento porcentual: (Vivus, 2022)

$$\text{Incremento porcentual} = (\text{Valor final} - \text{Valor inicial}) / \text{Valor inicial} * 100$$

Reemplazando la formula se obtiene:

$$\text{Incremento porcentual } 2020-2021 = (2021-2022) / 2021 * 100$$

Esta fórmula se aplicará respectivamente en todos los años desde el 2020 hasta la predicción del 2025 para que al final se cojan todos los incrementos porcentuales y sacarles un promedio, el resultado del promedio



indicará la media en que esa zona de la ciudad ha crecido en números de emergencias.

### Figura 27

Tabla con las zonas que presentaron mayor crecimiento en la delincuencia en los datos regresión lineal

|                                   |        |        |        |       |      |       |
|-----------------------------------|--------|--------|--------|-------|------|-------|
| GUAYAQUIL - AREA DE EXPANSION PRO | 111,20 | -7,78  | -1,54  | -1,21 | 1,19 | 20,37 |
| GUAYAQUIL-ATARAZANA               | 80,91  | -1,75  | -12,72 | -0,71 | 0,70 | 13,29 |
| GUAYAQUIL-MONTE BELLO             | 64,02  | -7,20  | 3,46   | -0,81 | 0,80 | 12,05 |
| GUAYAQUIL-LOS CEIBOS              | 62,76  | 2,85   | -6,04  | -0,80 | 0,80 | 11,91 |
| GUAYAQUIL- TRINITARIA 2           | 69,12  | -24,88 | 8,96   | -0,57 | 0,56 | 10,64 |
| GUAYAQUIL- CHILE                  | 64,97  | -12,46 | -1,74  | -0,49 | 0,48 | 10,15 |
| GUAYAQUIL- VICTORIA               | 62,43  | -8,25  | -3,57  | -0,51 | 0,50 | 10,12 |
| GUAYAQUIL-PUERTO HONDO            | 31,65  | 24,77  | -7,66  | -1,07 | 1,07 | 9,75  |
| GUAYAQUIL-ALBORADA                | 62,05  | -5,11  | -8,96  | -0,43 | 0,42 | 9,59  |
| GUAYAQUIL-MALVINAS 2              | 58,45  | -20,72 | 9,08   | -0,53 | 0,52 | 9,36  |

Nota: En esta tabla realizada por el autor se observa el crecimiento de las emergencias cada año junto a su promedio.

### Figura 28

Tabla con las zonas que presentaron mayor crecimiento en la delincuencia en los datos random forest

|                                   |        |        |        |       |       |       |
|-----------------------------------|--------|--------|--------|-------|-------|-------|
| GUAYAQUIL - AREA DE EXPANSION PRO | 111,20 | -7,21  | -7,91  | 3,55  | 0,27  | 19,98 |
| GUAYAQUIL-ATARAZANA               | 80,91  | -1,72  | -7,86  | 0,47  | -0,35 | 14,29 |
| GUAYAQUIL-URDESA                  | 54,16  | 20,91  | -6,75  | 2,13  | 0,11  | 14,11 |
| GUAYAQUIL-LOS CEIBOS              | 62,76  | 2,93   | -1,61  | 3,96  | -1,01 | 13,41 |
| GUAYAQUIL-MONTE BELLO             | 64,02  | -6,71  | 2,29   | 4,71  | 0,86  | 13,03 |
| GUAYAQUIL-SAN FRANCISCO           | 32,85  | 38,78  | -13,90 | 6,17  | -0,70 | 12,64 |
| GUAYAQUIL-PUERTO HONDO            | 31,65  | 32,93  | -4,23  | -0,96 | 0,41  | 11,96 |
| GUAYAQUIL-ALBORADA                | 62,05  | -4,86  | -0,89  | 2,47  | -0,50 | 11,65 |
| GUAYAQUIL-PUENTE LUCIA            | -12,53 | -6,76  | 44,38  | 20,10 | 12,15 | 11,47 |
| GUAYAQUIL- CHILE                  | 64,97  | -11,08 | 4,16   | -0,99 | -0,56 | 11,30 |

Nota: En esta tabla realizada por el autor se observa el crecimiento de las emergencias cada año junto a su promedio.

Con la comparación de ambos modelos de predicción con minería de datos se observa como la regresión lineal muestra un crecimiento mayor en su promedio en comparación que el modelo de random forest pero aun así ambas tablas comparten muchas de las zonas en común a excepción de Trinitaria 2, Victoria, Malvinas 2, Urdesa, San Francisco y Puente Lucia.

## **Comparación de rendimiento de los modelos**

Como una de las últimas etapas del diseño de un modelo predictivo usando minería de datos es la comparación y análisis de los resultados arrojados por las corridas ejecutadas tras el entrenamiento y prueba de los algoritmos de machine learning usados.

En esta parte se comparara específicamente los valores que se obtuvieron de la métrica de regresión llamada  $r^2\_score$ , como ya se mencionó anteriormente, esta variable métrica contiene un valor entre 0 y 1 en donde mientras más se acerca al 1 mejor se ajusta el modelo de regresión con los datos reales y mientras más se acerca a 0 representa un peor ajuste. (IBM, IBM, 2023)

En las corridas del algoritmo de regresión lineal en donde en el primer escenario se escogió como variable de entrada el año 2021 y variable de salida fue de 2022, se escogió esta relación de entradas y salidas porque se probó que en regresión lineal no es necesario tener tantas variables de entrada para tener una predicción ya que en ambas ocasiones las predicciones eran casi las mismas. Para el segundo escenario se uso solamente una variable de entrada siendo esta el año 2022 para predecir el valor del año 2024. Con la predicción del año 2025 fue parecido al solo usar el año 2023 como entrada.

**Tabla 8**

*Comparación de los algoritmos de regresión lineal, random forest y SVR*

|                  |          | r2_score | MAE       | MSE           |
|------------------|----------|----------|-----------|---------------|
| Regresion lineal | Año 2023 | 0,8911   | 760,4109  | 1282700,2819  |
|                  | Año 2024 | 0,8912   | 724,3702  | 1142230,0531  |
|                  | Año 2025 | 0,9998   | 28,1511   | 1273,9257     |
| Random Forest    | Año 2023 | 0,9798   | 360,4107  | 237298,2698   |
|                  | Año 2024 | 0,9984   | 95,1171   | 16793,9046    |
|                  | Año 2025 | 0,9993   | 56,0356   | 6413,0883     |
| SVR              | Año 2023 | -0,0058  | 2666,3193 | 11847208,9601 |

*Nota: En esta tabla se presentan los valores de r2, MAE y MSE en donde se comparan entre ellos y entre los algoritmos usados, tabla hecho por autor*

Con estos resultados se puede concluir que el algoritmo con menos precisión en estas corridas es el algoritmo de Regresión Lineal, pero aun así eso no signifique que las predicciones hechas con este algoritmo sean erróneas, simplemente presentan una precisión inferior que Random Forest. Esto se puede notar claramente en el valor de r2\_score en donde no hay ningún escenario en que la precisión baje de 0.8, el cual ya es en si un valor bastante cercano al 1.

En los valores de MAE se puede ver también como el algoritmo de random forest es más preciso que la regresión lineal ya que el error medio absoluto calcula como la media de los errores absolutos entre los valores reales y los valores predichos, dando a entender que mientras más bajo sea el valor de MAE, más preciso es el modelo. De igual manera aplica para MSE que es el error medio cuadrático el cual calcula como la media de los errores al cuadrado entre los valores reales y los valores predichos.

Finalmente se agregó a la tabla comparativa también la predicción hecha con el algoritmo de soporte de máquinas vectoriales, el cual no dio una predicción para nada precisa ya que con solo observar el  $r^2\_score$  su valor no se encuentra ni si quiera en los valores de los positivos por lo cual se descarta este algoritmo para este caso en específico.

Con esto se puede concluir que, de los tres algoritmos probados para la predicción de las emergencias en los años siguientes, los algoritmos de regresión lineal y random forest son opciones muy viables si se desea realizar una predicción con una precisión bastante alta para modelos de regresión usando un modelo de minería de datos a diferencia de SVR el cual no presento resultados deseados.

## ANÁLISIS DE LAS PREDICCIONES OBTENIDAS

**Figura 29**

Dashbord elaborado en Power BI que contiene las predicciones hechas del modelo de minería de datos

| Tabla de datos de algoritmo de Regresión Lineal |               |               |               |               |               |                   |                       |  |
|---|---------------|---------------|---------------|---------------|---------------|-------------------|-----------------------|--|
| DPA_DESP  | 2020          | 2021          | 2022          | 2023          | 2024          | 2025              | Crecimiento 2020-2021 |  |
| GUAYAQUIL-VENEZUELA                             | 7762          | 9822          | 8261          | 8605          | 8632          | 8.604,43          | 26,54                 |  |
| GUAYAQUIL-URDESA                                | 3918          | 6040          | 7303          | 5519          | 5494          | 5.518,77          | 54,16                 |  |
| GUAYAQUIL-UNION DE BANANEROS                    | 6256          | 8141          | 6091          | 7225          | 7237          | 7.224,58          | 30,13                 |  |
| GUAYAQUIL-SUBURBIO                              | 8553          | 13349         | 10762         | 11595         | 11560         | 11.594,09         | 56,07                 |  |
| GUAYAQUIL-SAUCES                                | 8746          | 13010         | 11527         | 11297         | 11278         | 11.296,13         | 48,75                 |  |
| GUAYAQUIL-SAN FRANCISCO                         | 3224          | 4283          | 5944          | 4041          | 4035          | 4.040,94          | 32,85                 |  |
| GUAYAQUIL-SAMANES                               | 3288          | 4392          | 4143          | 4132          | 4126          | 4.131,93          | 33,58                 |  |
| GUAYAQUIL-SALINAS                               | 9979          | 12567         | 11794         | 10868         | 10911         | 10.867,18         | 25,93                 |  |
| GUAYAQUIL-PUERTO HONDO                          | 932           | 1227          | 1631          | 1515          | 1499          | 1.515,22          | 31,65                 |  |
| GUAYAQUIL-PUENTE LUCIA                          | 862           | 754           | 703           | 1113          | 1107          | 1.113,27          | -12,53                |  |
| GUAYAQUIL-PORTETE                               | 12468         | 16312         | 15711         | 13977         | 14019         | 13.975,83         | 30,83                 |  |
| GUAYAQUIL-PASCUALES                             | 4507          | 6896          | 6152          | 6229          | 6204          | 6.228,69          | 53,01                 |  |
| GUAYAQUIL-BARRIO DE LA FLOR                     | 1410          | 5507          | 4407          | 5000          | 5070          | 5.007,00          | 35,00                 |  |
| <b>Total</b>                                    | <b>320850</b> | <b>440505</b> | <b>391130</b> | <b>391103</b> | <b>391080</b> | <b>391.080,00</b> | <b>1.976,19</b>       |  |

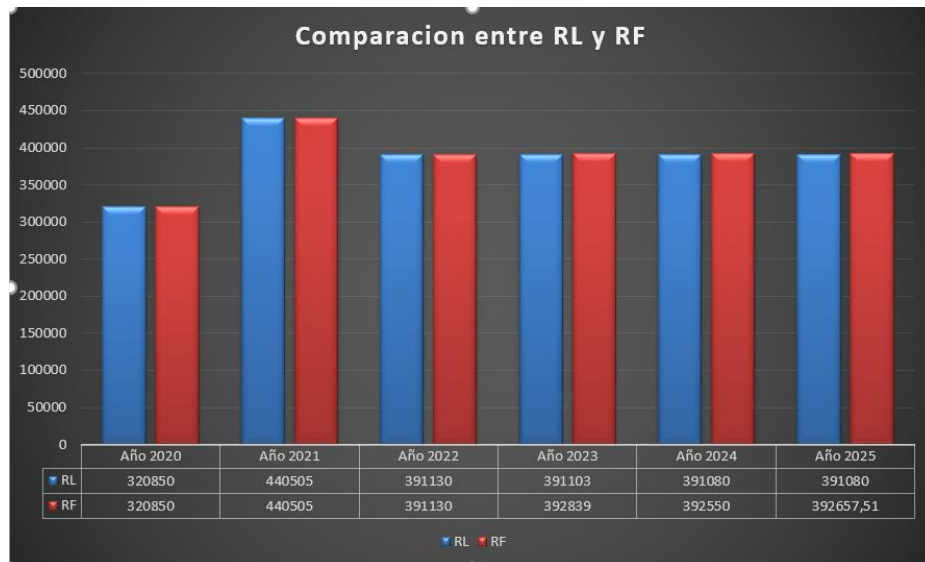
  

| Tabla de datos de algoritmo de Random Forest |               |               |               |               |               |                   |                               |                            |
|--|---------------|---------------|---------------|---------------|---------------|-------------------|-------------------------------|----------------------------|
| DPA_DESP                                     | 2020          | 2021          | 2022          | 2023          | 2024          | 2025              | Suma de Crecimiento 2020-2021 | Suma de Crecimie 2021-2022 |
| GUAYAQUIL - AREA DE EXPANSION PROGRESO       | 3124          | 6598          | 6122          | 5638          | 5838          | 5.854,03          | 111,20                        | -7                         |
| GUAYAQUIL-ATARAZANA                          | 7427          | 13436         | 13205         | 12167         | 12224         | 12.181,56         | 80,91                         | -1                         |
| GUAYAQUIL-URDESA                             | 3918          | 6040          | 7303          | 6810          | 6955          | 6.962,37          | 54,16                         | 20                         |
| GUAYAQUIL-LOS CEIBOS                         | 2221          | 3615          | 3721          | 3661          | 3806          | 3.767,40          | 62,76                         | 2                          |
| GUAYAQUIL-MONTE BELLO                        | 2279          | 3738          | 3487          | 3567          | 3735          | 3.767,06          | 64,02                         | -6                         |
| GUAYAQUIL-SAN FRANCISCO                      | 3224          | 4283          | 5944          | 5118          | 5434          | 5.395,78          | 32,85                         | 38                         |
| GUAYAQUIL-PUERTO HONDO                       | 932           | 1227          | 1631          | 1562          | 1547          | 1.553,32          | 31,65                         | 32                         |
| GUAYAQUIL-ALBORADA                           | 7537          | 12214         | 11620         | 11517         | 11801         | 11.742,01         | 62,05                         | -4                         |
| GUAYAQUIL-PUENTE LUCIA                       | 862           | 754           | 703           | 1015          | 1219          | 1.367,13          | -12,53                        | -6                         |
| GUAYAQUIL- CHILE                             | 7317          | 12071         | 10734         | 11180         | 11069         | 11.006,89         | 64,97                         | -11                        |
| <b>Total</b>                                 | <b>320850</b> | <b>440505</b> | <b>391130</b> | <b>392839</b> | <b>392550</b> | <b>392.657,51</b> | <b>1.976,19</b>               | <b>-471</b>                |

Nota: En este dashbord se observa dos secciones, una perteneciente a los datos obtenidos de la regresión lineal y random forest representado en una tabla, gráficos hechos por autor.

**Figura 30**

Dashboard de barras que muestra la comparación de las predicciones hechas



Comparación entre la regresión lineal y random forest a través de los años, grafico realizado por estudiante.

Con los gráficos mostrados, se puede observar una comparación en grafico circular entre los algoritmos de regresión lineal y random forest en que se muestra la suma de cada año y su representación porcentual dando la impresión que fuera el mismo gráfico, pero no lo es ya que, si se diferencia una variación de los porcentajes entre gráficos al tener predicciones muy parecidas, pero no iguales. También se puede observar una tabla perteneciente a las emergencias de reportadas por la Ecu911 del 2020 hasta 2022 junto a las predicciones realizadas con ambos algoritmos de regresión hasta el 2025.

En las tablas mostradas aparte de mostrar las predicciones obtenidas, también se tiene valores adicionales como la tasa de crecimiento año a año que las presentaban las predicciones mostrando los años que han tenido mayor o menor aumento en su número de emergencias. Otro punto importante que se debe notar en ambas tablas es la última columna la cual presenta un

promedio de todas las tasas de crecimiento indicando que sector ha sido el cual ha mostrado o puede llegar a mostrar un mayor índice de peligro y delincuencia según las emergencias realizadas.

## CONCLUSIONES

Para cumplir el primer objetivo, se extrajo la información de las zonas de la ciudad de Guayaquil que han realizado un llamado de emergencia o auxilio a través del sistema de la Ecu911. Toda esa información está disponible en un portal web propio de la Ecu911 en donde muestran con gráficos y estadísticas las emergencias realizadas del periodo entre el 2020 y la actualidad, por lo cual se decidió escoger el periodo del 1 de enero del 2020 hasta el 31 de diciembre del 2022. La información se la separo en cada año para no tener un solo registro de emergencias que conste de tres años. En el portal web uno puede filtrar los datos que desea visualizar como el canto, provincia y hasta parroquia por lo cual se filtró para que solo muestre las parroquias urbanas de Guayaquil. Tras haber filtrado la información que se iba a extraer ahora se tuvo que separar las variables o tipos de emergencias que tengan relación con la delincuencia en la ciudad como por ejemplo robo de autos, hurto, tráfico de drogas, entre muchas otras más y separar también variables que no se relacionaban en nada con la delincuencia como accidentes de tránsito o catástrofes naturales. Tras filtrar las variables que, si tenían relación con la delincuencia, se procedió a descargar el set de datos para posteriormente usarla en el proceso de minería de datos.

Para cumplir el segundo objetivo de elaborar un modelo predictivo con minería de datos para calificar a los sectores que se pueden volver peligroso se optó por sacar la probabilidad de crecimiento porcentual de cada año desde el 2020 hasta el 2025 con una formula la cual al valor futuro de la predicción se le resta el valor del año anterior para esa resta dividirla por el valor del año anterior y multiplicarlo por 100. Realizando este proceso se obtuvo el



crecimiento o disminución porcentual de cada año para a continuación calcular un promedio de todos los crecimientos y disminuciones porcentuales, obteniendo como resultado un valor que indica el promedio de los sectores que tienden a tener un mayor crecimiento tras realizar las predicciones.

Por último, para cumplir el objetivo de valorar la aproximación de las predicciones generadas a través del modelo predictivo se ejecutó por igual en los códigos de los algoritmos de regresión lineal, random forest y SVR o también conocido como máquinas de soporte de vectores en donde se importaba la librería panda, numpy y sklearn que era la librería que aportaba los métodos de aprendizaje supervisados como son las regresiones. La librería scikit learn también ofrece métodos que ayudan a comprobar que tan acertado o erróneo se crea el modelo con los paquetes de `mean_squared_error`, `r2_score` y `mean_absolute_error` los cuales corresponden a

- Error cuadrático medio (MSE) que calcula como la media de los errores al cuadrado entre los valores reales y los valores predichos
- `R2_score` se define como la proporción de la varianza en los valores reales explicada por la varianza en los valores predichos. Un `R2` cercano a 1 indica una buena precisión de la predicción, mientras que un `R2` cercano a 0 indica una precisión baja.
- Error absoluto medio (MAE) el cual calcula como la media de los errores absolutos entre los valores reales y los valores predichos.

De estos tres valores de que miden la aproximación y error del modelo predictivo el que más impacto tiene es el `r2_score` ya que como se menciona mientras más cercano al 1 se encuentre su valor más

aproximado es el dato predicho y en este caso de los tres algoritmos que se implementaron, el único que no presento aproximación alguna al mostrar valores en los rangos de los negativos fue el soporte de máquinas vectoriales, mientras que la regresión lineal y random forest tenían una aproximación `r2_score` de más del 0.8 lo cual podría representarse como un 80% de acertamiento del modelo predictivo.

## RECOMENDACIONES

Como recomendación para futuras investigaciones de un mismo ámbito de desarrollo se podría recomendar lo siguiente:

- Obtener un dataset más amplio que permita tener varios datos históricos para que la predicción sea más exacta.
- Obtener datos de diferentes fuentes ya sean obtenidas de servicios públicos o recopiladas a través de investigación.
- Se recomendaría usar más algoritmos de machine learning con un set de datos más amplios para determinar diferentes predicciones.
- Se podría usar otras herramientas de desarrollo como lo es Google Colab el cual presenta una interfaz cómoda para el usuario con las mismas funciones que Jupyter Notebook.
- Se podría desarrollar en otros lenguajes de programación adecuados para la minería de datos como lo es R y Java.

## REFERENCIAS

Amazon. (s.f). *aws.amazon.com*. <https://aws.amazon.com/es/what-is/linear-regression/#:~:text=La%20regresi%C3%B3n%20lineal%20es%20una,independiente%20como%20una%20ecuaci%C3%B3n%20lineal>.

Ammar, A. (3 de Junio de 2022). *astera.com*. <https://www.astera.com/es/tipo/blog/Las-10-mejores-t%C3%A9cnicas-de-miner%C3%ADa-de-datos/>

Andalucía. (23 de Marzo de 2022). *juntadeandalucia.es*. <https://www.juntadeandalucia.es/datosabiertos/portal/actualidad/detalle/1228>

Bello, E. (20 de Diciembre de 2021). *iebschool*. <https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/>

BrianBlanchard, a. v.-r.-t.-r. (22 de Agosto de 2022). *learn.microsoft.com*. <https://learn.microsoft.com/es-es/azure/cloud-adoption-framework/innovate/considerations/predict>

Coffey, H. (23 de Julio de 2022). *Independent en Español*. <https://www.independentespanol.com/noticias/paises-peligrosos-mundo-cuales-son-b2129889.html>

Cordoba, L. (16 de Junio de 2011). <http://cor-mineriadedatos.blogspot.com/2011/06/weka.html#:~:text=Weka%20co>

ntiene%20una%20colecci%C3%B3n%20de,est%C3%A9n%20a%20una%20mejor%20disposici%C3%B3n.

Domínguez, V. H., Sosa, J. D., Bolaños, M. E., & Pérez, J. W. (2022). Uso de la minería de datos para la caracterización de investigadores y cuerpos académicos. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, XII(24).

Edx. (25 de Marzo de 2021). *blog.edx.org*. <https://blog.edx.org/es/r-vs-python-para-la-ciencia-de-datos-explicacion-y-consejos-de-aprendizaje#:~:text=3.,necesarios%20para%20ponerlos%20en%20marcha>.

España, S. (19 de Octubre de 2021). *elPais*. <https://elpais.com/internacional/2021-10-20/ecuador-el-pais-donde-las-balas-no-distinguen-barrios-ni-horarios.html>

Espinosa, J. (2 de Diciembre de 2020). *scielo.org*. [https://www.scielo.org.mx/scielo.php?script=sci\\_arttext&pid=S1405-77432020000300002#:~:text=El%20algoritmo%20Random%20Forest%20\(Breiman,de%20cada%20%C3%A1rbol%20por%20separado%20\(](https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432020000300002#:~:text=El%20algoritmo%20Random%20Forest%20(Breiman,de%20cada%20%C3%A1rbol%20por%20separado%20()

Galán, V. (Octubre de 2015). *Universidad Carlos III de Madrid*. [https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC\\_Victor\\_Galan\\_Cortina.pdf](https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf)

IBM. (27 de Febrero de 2021). *IBM*.  
<https://www.ibm.com/docs/es/db2/11.1?topic=miner-data-mining-process>

IBM. (17 de Agosto de 2021). *ibm.com*. <https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-about-svm>

IBM. (3 de Enero de 2023). *IBM*. <https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=terms-r2>

INEC. (2010). *INEC*. <https://www.ecuadorencifras.gob.ec/estadisticas/>

inesdi. (25 de Agosto de 2021). *inesdi*. Digital Business School:  
<https://www.inesdi.com/blog/herramientas-y-tecnicas-de-data-mining/>

inesdi. (3 de Junio de 2021). *inesdi*. Digital Business School:  
<https://www.inesdi.com/blog/que-es-el-data-mining/>

IONOS. (28 de Febrero de 2019). *ionos.es*.  
<https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/jupyter-notebook/>

IONOS. (10 de Mayo de 2022). *ionos.es*. Analisis web:  
<https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/>

Jupyter. (s.f). *jupyter.org*. <https://jupyter.org/>

KeepCoding. (24 de Octubre de 2022). *keepcoding.io*.  
<https://keepcoding.io/blog/que-es-random->



- Mundo, B. N. (2 de Noviembre de 2022). *BBC*.  
<https://www.bbc.com/mundo/noticias-america-latina-63487343>
- Otzen, T., & Monterola, C. (2017). *scielo.cl*.  
[https://www.scielo.cl/scielo.php?script=sci\\_arttext&pid=S0717-95022017000100037](https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-95022017000100037)
- Pichel, M. (11 de Octubre de 2021). *BBC News Mundo* .  
<https://www.bbc.com/mundo/noticias-america-latina-58829554>
- PlanV. (3 de Septiembre de 2018). *PlanV*.  
<https://www.planv.com.ec/historias/sociedad/guayaquil-la-ciudad-con-mas-incautaciones-droga-ecuador>
- Román, A. L. (6 de Noviembre de 2022). *eltiempo*.  
<https://www.eltiempo.com/mundo/latinoamerica/violencia-en-ecuador-los-crimenes-de-las-carceles-se-trasladan-a-calles-715456>
- Roman, V. (25 de Abril de 2019). *medium.com*. <https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f>
- Sanchez, A. (14 de Junio de 2022).  
<https://aprendeconalf.es/docencia/python/manual/pandas/>
- Santander, U. (10 de Diciembre de 2021). *Santander*. <https://www.becas-santander.com/es/blog/cualitativa-y-cuantitativa.html#:~:text=Los%20datos%20recogidos%20se%20pued en,las%20palabras%20y%20los%20significados.>



Sotaquirá, M. (14 de Marzo de 2021). *codificandobits.com*.

<https://www.codificandobits.com/blog/maquinas-de-soporte-vectorial/#:~:text=El%20principal%20uso%20de%20las,dos%20categorias%C3%ADas%20o%20clases%20diferentes>.

Unir. (22 de Mayo de 2020). *unir.net*. La Universidad en Internet:

<https://www.unir.net/ingenieria/revista/analisis-predictivo/>

Universia. (17 de Diciembre de 2020). *universia.net*.

<https://www.universia.net/es/actualidad/orientacion-academica/para-que-sirve-phyton-que-es-y-usos-1154393.html>

Veigler. (23 de Abril de 2021). *veigler Business School*.

<https://veigler.com/data-mining/>

Vivus. (11 de Marzo de 2022). *vivus.es*. <https://www.vivus.es/blog/como-calcular-porcentaje>

Zambrano, J.-C., Quieroz, P., Santamaria, A., & Zamora, W. (2022). Covid-19

en Ecuador: Aplicación de minería de datos. *Informatica y Sistemas*,

VI(1), 35-52. <http://revistas.utm.edu.ec/index.php/informaticaysistemas>

# ANEXOS

|                                     |                           |               |         |
|-------------------------------------|---------------------------|---------------|---------|
| <input type="checkbox"/>            | VF Random forest.ipynb    | hace 6 horas  | 32.3 kB |
| <input type="checkbox"/>            | VF Regresion lineal.ipynb | hace 8 horas  | 192 kB  |
| <input type="checkbox"/>            | VF RF 2023.ipynb          | hace 6 horas  | 34.8 kB |
| <input type="checkbox"/>            | VF RF 2024.ipynb          | hace 6 horas  | 36.9 kB |
| <input type="checkbox"/>            | VF RL 2020 y 2021.ipynb   | hace 8 horas  | 365 kB  |
| <input type="checkbox"/>            | VF RL 2023.ipynb          | hace 7 horas  | 520 kB  |
| <input type="checkbox"/>            | VF RL2024.ipynb           | hace 6 horas  | 37.3 kB |
| <input type="checkbox"/>            | VF SVR.ipynb              | hace 13 horas | 2.81 kB |
| <input type="checkbox"/>            | datos2020-22-gpt.csv      | hace 7 días   | 2.17 kB |
| <input type="checkbox"/>            | Datos2020-22.csv          | hace 18 días  | 3.02 kB |
| <input type="checkbox"/>            | datos2020.csv             | hace 2 días   | 886 B   |
| <input type="checkbox"/>            | datos2023rf.csv           | hace 6 horas  | 3.26 kB |
| <input type="checkbox"/>            | datos2024rf.csv           | hace 6 horas  | 3.56 kB |
| <input type="checkbox"/>            | datos2024RL.csv           | hace 7 horas  | 3.56 kB |
| <input checked="" type="checkbox"/> | DatosRL.csv               | hace 11 horas | 3.31 kB |

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor

df = pd.read_csv("Datos2020-22.csv")

# Preparar los datos para el modelo
X = df[['StatusIncident2020', 'StatusIncident2021']]
y = df['StatusIncident2022']

# Entrenar el modelo
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X, y)
```

```
# Hacer la predicción para 2023
poblacion_2023 = model.predict(X)
print("La población prevista para 2023 es:", poblacion_2023)
```

```
La población prevista para 2023 es: [14729.75 13062.02 14874.02 11349.82 12097.49 10592.33 13570.35 10509.07
11683.23 8052.97 7956.16 11322.86 8268.25 11631.59 12237.68 11408.83
8153.78 8254.48 6025. 7486.87 8183.06 8923.86 6497.08 6629.39
7485.03 6105.03 7253.46 5805.64 8461.53 5517.88 5016.29 4460.89
7543.88 7266.62 6806.6 6446.98 6322.42 4723.23 5031.07 5663.21
5944.33 6779.83 3994.23 3651.3 4557.44 5120.82 6151.67 3482.68
3461.36 2896.73 1537.35 1476.27 1015.32]
```

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)
```

```
MSE: 237298.26982452828
MAE: 360.41075471698105
r2: 0.9798535172219616
```

```
df.insert(6, 'StatusIncident2023', poblacion_2023)
```

```
df
```

|   | DPA_DESPRO | DPA_DESCAN | DPA_DESP                   | StatusIncident2020 | StatusIncident2021 | StatusIncident2022 | StatusIncident2023 |
|---|------------|------------|----------------------------|--------------------|--------------------|--------------------|--------------------|
| 0 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-PORTETE          | 12468              | 16312              | 15711              | 14596.02           |
| 1 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL- EL CISNE        | 12086              | 15839              | 12877              | 12407.22           |
| 2 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-GARAY            | 11932              | 16989              | 15298              | 15069.60           |
| 3 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-GUASMO           | 10897              | 13840              | 9999               | 10938.52           |
| 4 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-SALINAS          | 9979               | 12567              | 11794              | 12405.72           |
| 5 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-NUEVA PROSPERINA | 9366               | 11495              | 10107              | 10162.46           |
| 6 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-MARTHA ROLDOS    | 9157               | 12724              | 14975              | 13829.73           |
| 7 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL- Esteros         | 8939               | 15830              | 9859               | 11018.94           |
| 8 | GUAYAS     | GUAYAQUIL  | GUAYAQUIL-SAUCES           | 8746               | 13010              | 11527              | 12484.90           |

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor

df = pd.read_csv("datos2023rf.csv")

# Preparar los datos para el modelo
X = df[['StatusIncident2020', 'StatusIncident2021', 'StatusIncident2022']]
y = df['StatusIncident2023']

# Entrenar el modelo
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X, y)

# Hacer la predicción para 2024
poblacion_2024 = model.predict(X)
print("La población prevista para 2024 es:", poblacion_2024)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2024)
mae = mean_absolute_error(y, poblacion_2024)
r2 = r2_score(y, poblacion_2024)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)
```

```
La población prevista para 2024 es: [14538.97 12479.91 14610.6  10972.03 12173.16 10358.92 13753.3  11167.9
12178.97 8026.56 7963.05 11281.28 8230.19 11731.31 12128.46 11117.19
8173.89 8322.3 6037. 7322.93 8077.9 8687.04 6499.34 6822.81
7517.42 5959.9 7260.13 5769.78 8293.71 5799.59 5158.4 4556.51
7365.04 7416.14 6855.74 6149.27 6318.4 4778.33 5066.27 5461.57
5880.87 6844.16 4119.83 3702.59 4456.77 5270.84 5831.65 3584.09
3586.79 3084.88 1536.54 1515.72 1221.09]
MSE: 16793.9046245283
MAE: 95.11716981132079
r2: 0.9984775307239794
```

```

import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor

df = pd.read_csv("datos2024rf.csv")

# Preparar los datos para el modelo
X = df[['StatusIncident2020', 'StatusIncident2021', 'StatusIncident2022', 'StatusIncident2023']]
y = df['StatusIncident2024']

# Entrenar el modelo
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X, y)

# Hacer la predicción para 2025
poblacion_2025 = model.predict(X)
print("La población prevista para 2025 es:", poblacion_2025)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2025)
mae = mean_absolute_error(y, poblacion_2025)
r2 = r2_score(y, poblacion_2025)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

```

La población prevista para 2025 es: [14338.27 12311.71 14402.15 10800.05 12105.47 10564.48 13809.22 11022.67
12157.66 7928.92 7860.2 11194.08 8113.27 11742.01 12181.56 11006.89
7988.67 8107.5 5941.49 7236.68 8209.97 8594.27 6286.72 6892.92
7525.37 5923.4 7218.89 5796.89 8335.8 5818.93 5152.89 4188.28
7477.93 7719.66 6932.1 6063.87 6268.43 4768.92 5115.87 5477.79
6008.96 6962.37 4040.89 3937.67 4564.59 5395.78 5854.03 3767.06
3767.4 3298.25 1558.21 1553.32 1367.13]
MSE: 6413.088307547182
MAE: 56.03566037735853
r2: 0.9993969715823626

```

```

import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression

df = pd.read_csv("Datos2020-22.csv")

# Preparar los datos para el modelo
X = df[['StatusIncident2020', 'StatusIncident2021']]
y = df['StatusIncident2022']

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

```

```
LinearRegression()
```

```

# Hacer la predicción para 2023
poblacion_2023 = model.predict(X)
print("La población prevista para 2023 es:", poblacion_2023)

```

```

La población prevista para 2023 es: [13977.08768572 13587.16430467 14578.72567522 11922.15526659
10868.31461981 9974.43308254 11035.89686706 13706.01455416
11297.48265248 8512.96370205 7993.14896219 11595.71381869
8605.44436626 10664.05856507 11715.54300816 10550.37786067
8242.8382448 9128.09204871 6341.86155843 7644.92010801
7748.83030796 9486.28303771 7225.65968346 6815.95638815
7152.2159344 6678.82483346 7192.9086367 6470.55417116
8302.04277657 5594.66416688 5207.3592814 5910.21081107
7459.58176671 6373.69739386 6143.70519759 6402.19656939
6229.23064894 5068.32178542 5055.27475783 5737.41792368
5401.40301832 5519.46172847 3980.64317027 3702.99291056
4132.72303012 4041.90114624 6029.51144845 3612.92713855
3509.8681703 2658.39648525 1714.33098872 1515.56935985
1113.09838126]

```

```

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

MSE: 1281428.7283904164  
MAE: 752.0649487929595  
r2: 0.891207458752686

```

# Preparar los datos para el modelo
X = df[['StatusIncident2020']]
y = df['StatusIncident2022']

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

# Hacer la predicción para 2023
poblacion_2023 = model.predict(round(X, 0))
print("La población prevista para 2023 es:", poblacion_2023)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

La población prevista para 2023 es: [14482.76080858 14059.7438259 13889.20818367 12743.07578295  
11726.50617535 11047.68572932 10816.24450057 10574.83690313  
10361.11366319 10307.95969678 10158.46416625 10147.39042325  
9271.45735178 9022.29813423 8900.4869612 8778.67578818  
8733.27344187 8682.33422406 8015.69489533 7950.35981162  
7849.5887503 7737.74394598 7603.75165565 7583.81891825  
7526.23545464 7514.05433734 7396.67266151 7326.9080806  
7268.21724269 7261.57299689 6388.96204832 6382.31780252  
6376.78093102 6336.91545621 6276.0098697 6142.01757937  
5666.95400458 5569.50506616 5433.29802724 5076.72350257  
5056.79076516 5014.71054175 4682.49825169 4541.86171556  
4317.06473262 4246.19277741 4135.45534738 3199.7240637  
3135.49635429 3105.59724819 1968.32384187 1708.09088132  
1630.5746803 ]

MSE: 2435935.7782707713  
MAE: 1241.802881297607  
r2: 0.7931904929537456

```

# Preparar los datos para el modelo
X = df[['StatusIncident2021']]
y = df['StatusIncident2022']

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

# Hacer la predicción para 2023
poblacion_2023 = model.predict(round(X, 0))
print("La población prevista para 2023 es:", poblacion_2023)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)
sns.scatterplot(x='StatusIncident2021', y='DPA_DESP', data=df)
sns.scatterplot(x=poblacion_2023, y='DPA_DESP', data=df)

```

```

La población prevista para 2023 es: [14019.59890134 13627.05016114 14581.44984663 11968.05453393
10911.57557773 10021.90908831 11041.87188262 13619.58094621
11279.22693484 8580.35060684 8071.61407883 11560.56736387
8633.46502411 10618.61636993 11632.76977486 10499.93884382
8263.3239287 9118.96399456 6396.85010899 7656.65769384
7753.75748793 9432.67102161 7238.38165776 6840.85344094
7164.51942123 6705.57765943 7199.37575757 6497.26955416
8269.13331809 5646.60896492 5240.7816204 5921.31009179
7421.79237993 6368.63307481 6143.72671415 6389.38089406
6205.14025913 5077.28880471 5059.86063654 5708.0225099
5381.8667913 5494.73492802 3992.59281434 3718.72160024
4127.03868307 4036.57819115 5957.82625367 3584.2757315
3482.19646079 2656.43325465 1702.03356917 1500.36476606
1107.81602585]
MSE: 1282700.2819010115
MAE: 760.4109278196763
r2: 0.8910995046115897

```

```

# Preparar los datos para el modelo
X = df[['StatusIncident2020', 'StatusIncident2021', 'StatusIncident2022']]
y = df['StatusIncident2023']

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

# Hacer la predicción para 2023
poblacion_2024 = model.predict(round(X, 0))
print("La población prevista para 2024 es:", poblacion_2024)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2024)
mae = mean_absolute_error(y, poblacion_2024)
r2 = r2_score(y, poblacion_2024)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

La población prevista para 2024 es: [13976.53827521 13586.78462029 14578.31166889 11921.81984055  
10867.75709677 9973.9082047 11035.174866 13706.02134316  
11297.06171133 8512.43791792 7992.56660489 11595.39642204  
8604.94962185 10663.60695912 11715.12252596 10549.98536422  
8242.33207806 9127.69536913 6341.29308916 7644.44532628  
7748.2482679 9485.94618617 7225.21194972 6815.40379451  
7151.64333595 6678.32871903 7192.3803208 6470.05488077  
8301.57884679 5594.03316947 5206.79241247 5909.82505617  
7459.1029027 6373.04273979 6143.11014472 6401.70254433  
6228.73322728 5067.79569572 5054.71073621 5736.94867722  
5400.85114249 5518.81741656 3980.07977116 3702.39760321  
4132.15229373 4041.1929559 6029.05801623 3612.38436791  
3509.29793762 2657.79409953 1713.72919617 1514.94855059  
1112.49413565]
MSE: 0.07173760204615207
MAE: 0.22931260316518778
r2: 0.999999993166323

```

import numpy as np
from sklearn.linear_model import LinearRegression
df = pd.read_csv("DatosRL.csv")
df.head()

# Preparar los datos para el modelo
X = df[['StatusIncident2021', 'StatusIncident2022']]
y = df['StatusIncident2023']

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

# Hacer la predicción para 2023
poblacion_2024 = model.predict(round(X, 0))
print("La población prevista para 2024 es:", poblacion_2024)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2024)
mae = mean_absolute_error(y, poblacion_2024)
r2 = r2_score(y, poblacion_2024)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

```

La población prevista para 2024 es: [14020.78623438 13625.98012969 14581.75083023 11965.82687854
10911.95278071 10021.53318513 11045.06224684 13615.737326
11279.02714216 8579.2191923 8070.87070268 11559.40950818
8632.63862315 10619.0970492 11633.79750868 10499.71015517
8262.66957531 9117.66453465 6395.85322783 7655.57581196
7754.20680536 9431.47050314 7236.81225387 6840.30802021
7164.54670692 6704.23187238 7198.94420241 6495.9049636
8268.87171141 5646.25441626 5239.80673706 5918.58750623
7421.56315027 6369.79243648 6143.9628156 6388.60999401
6204.55735432 5076.16909168 5059.30199172 5707.02491255
5381.66769754 5495.85200501 3991.58744867 3717.98216709
4126.47606712 4037.75618354 5957.43838654 3583.597469
3481.82567882 2655.78707237 1701.28400332 1499.85297334
1106.80275937]
MSE: 1273.8191412217764
MAE: 27.974581803715594
r2: 0.9998786568236442

```



```

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

# Hacer la predicción para 2024
poblacion_2024 = model.predict(round(X, 0))
print("La población prevista para 2024 es:", poblacion_2024)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2024)
mae = mean_absolute_error(y, poblacion_2024)
r2 = r2_score(y, poblacion_2024)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

```

La población prevista para 2024 es: [14804.21074831 12278.49772899 14436.13753837 9713.57109893
11313.30817362 9809.82268894 14148.27398676 9588.80051929
11075.35285388 7823.29681737 7755.56421699 10393.57075798
8164.63347453 11158.2361675 12570.81737273 10368.61664205
8008.67024998 8129.87595591 6051.55458646 7066.65237368
8632.52314818 8498.84038428 6230.68949008 6875.04041209
7711.00329569 5982.93076765 7297.47794602 5783.29784023
8393.67661001 6019.47072312 5065.86700729 3967.88590644
7686.94039819 8112.94280582 7023.87388923 6263.66457185
6285.05381407 4783.35076624 5310.95207444 5450.87336733
5939.26106478 7310.84622241 3949.1703195 3967.88590644
4494.59599622 6099.68038146 6258.31726129 3909.95670875
4118.50182044 3132.81424127 2201.49098609 2255.85531008
1428.80461074]
MSE: 1142230.0531302511
MAE: 724.3702670095591
r2: 0.89119191391409

```

```

# Preparar los datos para el modelo
X = df[['StatusIncident2023']]
y = df['StatusIncident2024']

# Entrenar el modelo
model = LinearRegression()
model.fit(X, y)

# Hacer la predicción para 2025
poblacion_2025 = model.predict(X)
print("La población prevista para 2025 es:", poblacion_2025)

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2025)
mae = mean_absolute_error(y, poblacion_2025)
r2 = r2_score(y, poblacion_2025)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

```

La población prevista para 2025 es: [13975.8285546 13585.87214835 14576.76137551 11921.05826014
10867.17607506 9973.27600535 11034.15740799 13704.85884667
11296.12812193 8511.43942604 7992.49743926 11594.09481183
8604.4290306 10663.19887794 11714.08139837 10549.21162073
8241.46960633 9127.37057026 6340.68209793 7643.53645008
7747.52482508 9485.33055343 7224.58328542 6814.62911475
7151.59144528 6677.64442846 7191.58697413 6469.66767846
8301.4628996 5593.76559674 5206.80885515 5909.73027462
7458.55712917 6372.67852101 6142.70423014 6401.67527942
6228.69461716 5067.82439241 5054.82584554 5736.74961236
5400.78717005 5518.77398015 3979.94600781 3701.97708233
4131.92901742 4040.9391893 6028.71697293 3611.98714242
3508.99865565 2658.09377945 1714.19929869 1515.22154268
1113.26647778]
MSE: 1273.9257966577877
MAE: 28.151173072677246
r2: 0.9998786342611244

```

```

import pandas as pd
import numpy as np
from sklearn.svm import SVR

df = pd.read_csv("Datos2020-22.csv")

# Preparar Los datos para el modelo
X = df[['StatusIncident2020', 'StatusIncident2021']]
y = df['StatusIncident2022']

# Entrenar el modelo
model = SVR()
model.fit(X, y)

# Hacer La predicción para 2023
poblacion_2023 = model.predict(X)
print("La población prevista para 2023 es:", poblacion_2023)

```

```

La población prevista para 2023 es: [7034.81020184 7035.82541943 7034.54161413 7039.31157271 7040.68695842
7040.29829987 7041.22051197 7037.74413585 7041.21711772 7036.48173239
7034.45271154 7041.03786951 7035.6593723 7040.17147608 7040.03093095
7039.7546032 7033.31533462 7036.5448493 7023.34323602 7028.99999998
7029.2470127 7035.82567544 7026.23851485 7024.31984493 7025.71376455
7023.54650837 7025.59482437 7022.20773337 7030.39364253 7018.83591704
7015.95461734 7017.98936161 7024.51748888 7019.63756271 7018.62708692
7019.38010619 7017.91158555 7014.40382113 7014.22337523 7015.55072053
7014.62610006 7014.89077689 7012.68825355 7012.71458897 7012.65366724
7012.64180047 7015.93803614 7013.27560627 7013.39291017 7014.37483162
7017.2490489 7017.94078132 7018.8229855 ]

```

```

from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
mse = mean_squared_error(y, poblacion_2023)
mae = mean_absolute_error(y, poblacion_2023)
r2 = r2_score(y, poblacion_2023)
print("MSE:", mse)
print("MAE:", mae)
print("r2:", r2)

```

```

MSE: 11847208.960115278
MAE: 2666.3193505549093
r2: -0.005821034680439485

```

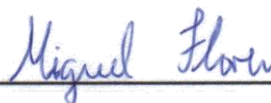
## DECLARACIÓN Y AUTORIZACIÓN

Yo, **Miguel Alfonso Flores Asinc**, con C.C: # **0924991912** autor/a del trabajo de titulación: “**Modelo predictivo que analice cuales son las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y a partir de esta información predecir qué sectores en desarrollo tendrán el mismo nivel de delincuencia.**” previo a la obtención del título de **Ingeniero en Sistemas Computacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para omsu difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, 16 de febrero del 2023



---

Nombre: **Flores Asinc Miguel Alfonso**

C.C: **0924991912**



## **REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA**

### **FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN**

|  |   |                                  |    |
|--|---|----------------------------------|----|
| <b>TEMA Y SUBTEMA:</b>   | Modelo predictivo que analice cuales son las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y a partir de esta información predecir qué sectores en desarrollo tendrán el mismo nivel de delincuencia   |                                  |    |
| <b>AUTOR(ES)</b>   | Miguel Alfonso Flores Asinc   |                                  |    |
| <b>REVISOR(ES)/TUTOR(ES)</b>                                       | Marcos Xavier Miranda Rodríguez   |                                  |    |
| <b>INSTITUCIÓN:</b>  | Universidad Católica de Santiago de Guayaquil   |                                  |    |
| <b>FACULTAD:</b>   | Ingeniería  |                                  |    |
| <b>CARRERA:</b>  | Ingeniería en Sistemas Computacionales  |                                  |    |
| <b>TITULO OBTENIDO:</b>  | Ingeniero en Sistemas Computacionales   |                                  |    |
| <b>FECHA DE PUBLICACIÓN:</b>                                       | 16 de febrero de 2023   | <b>No. DE PÁGINAS:</b>           | 89 |
| <b>ÁREAS TEMÁTICAS:</b>  | Modelos predictivos, Minería de datos   |                                  |    |
| <b>PALABRAS CLAVES/<br/>KEYWORDS:</b>                              | Minería de datos, Aprendizaje autónomo, Algoritmos de regresión, Modelo predictivo, KDD   |                                  |    |
| <b>RESUMEN/ABSTRACT:</b>   | <p>En el presente proyecto de investigación contiene el diseño de un modelo predictivo que analice las zonas con mayor índice de peligrosidad en la ciudad de Guayaquil y con esta información predecir qué sectores en desarrollo tendrán un aumento de delincuencia usando la minería de datos y algoritmos de aprendizaje de máquina. Este modelado se lo hará con el uso de herramientas de programación con las que se podrá limpiar los datos, procesarlos y aplicando aprendizaje supervisado junto con algoritmos de regresión y random forest y así de esta manera entrenar el modelo y poder predecir valores futuros de la delincuencia en las diferentes zonas de Guayaquil y poder ser visualizados en una herramienta de visualización de datos. Todo el proceso desde la recopilación de datos hasta el análisis de los resultados obtenidos forma parte de las etapas de la metodología de minería de datos KDD o Knowledge Discovery in Databases. Los objetivos del proyecto de investigación constan de recopilar y analizar variables que describan a un sector como peligroso, diseño propio del modelo predictivo y la valoración de las predicciones obtenidas tras la creación del modelo. Por último, el proyecto concluye con la visualización de los algoritmos de regresión en una herramienta de visualización que presente las zonas que puedan convertirse en peligrosas en un futuro.</p> |                                  |    |
| <b>ADJUNTO PDF:</b>  | <input checked="" type="checkbox"/> SI  | <input type="checkbox"/> NO      |    |
| <b>CONTACTO CON AUTOR/ES:</b>                                      | Teléfono: +593-981831194  | E-mail: mfloresasinc@hotmail.com |    |
| <b>CONTACTO CON LA INSTITUCIÓN (COORDINADOR DEL PROCESO UTE)::</b> | Toala Quimí, Edison José  |                                  |    |
|  | Teléfono: +593-990-976776   |                                  |    |
|  | E-mail: edison.toala@cu.ucsg.edu.ec   |                                  |    |
| <b>SECCIÓN PARA USO DE BIBLIOTECA</b>                              |   |                                  |    |
| <b>Nº. DE REGISTRO (en base a datos):</b>                          |   |                                  |    |
| <b>Nº. DE CLASIFICACIÓN:</b>                                       |   |                                  |    |
| <b>DIRECCIÓN URL (tesis en la web):</b>                            |   |                                  |    |